EPI BRIEFING PAPER
ECONOMIC POLICY INSTITUTE • OCTOBER 30, 2015 • EPI BRIEFING PAPER #410

BRINGING IT BACK HOME

Why state comparisons are more useful than international comparisons for improving U.S. education policy

BY **MARTIN CARNOY**, **EMMA GARCÍA**, AND **TATIANA KHAVENSON**

Table of contents

Introduction and key findings 3

Do U.S. students perform poorly on international tests? 6

Comparing U.S. student average performance on the 2012 PISA 7

Comparing the 2012 PISA performance of U.S. students with different levels of family academic resources 9

Comparing changes in the PISA performance of low FAR and high FAR students in the United States with their counterparts in other countries, 2000–2012 10

Comparing average 2011 TIMSS performance of students in U.S. states, the United States as a whole, and other countries 11

Changes in U.S. states' TIMSS scores over time 12

Why it is difficult to learn much about improving U.S. schools from international test comparisons 13

Why some of the policy recommendations are not feasible 14

Should U.S. policymakers look outward or inward? 16

The case for looking inward across states 16

What we can learn from student performance in U.S. states over time 17

Comparing states' performance on the NAEP over 1992–2013 using regression analysis 17

Adjusted score rankings for states, 2003–2013 18

Mathematics gains in states with low, middle, and high initial (2003) levels 19

Reading gains in states with low, middle, and high initial (2003) levels 20

State rankings by gain in 8th grade adjusted mathematics score over 1992–2013 21

Explaining variation in adjusted state performance 22

Matchups of states with recently different 8th grade student mathematics score trajectories 23

Conclusion 25

Acknowledgments 27

About the authors 27

Tables and figures 29

Appendix: Regression analysis of state scores adjusted for FAR measures 55

Endnotes 58

References 62

Introduction and key findings

Since its inception in 2000, the Program for International Student Assessment (PISA)[1]—an international test of reading, math, and science—has shown that American 15-year-olds perform more poorly, on average, than 15-year-olds in many other developed countries. This finding is generally consistent with results from another international assessment of 8th graders, the Trends in International Mathematics and Science Study (TIMSS).[2]

International test rankings have come to dominate how politicians and pundits judge the quality of countries' education systems, including highly heterogeneous systems such as that of the United States. While international tests and international comparisons are not without merit, international test data are notoriously limited in their ability to shed light on *why* students in any country have higher or lower test scores than in another.[3] Policy prescriptions based on these test results therefore risk being largely descriptive, based on correlational evidence that offers limited and less-than-convincing proof of the factors that actually drive student performance.

Indeed, from such tests, many policymakers and pundits have wrongly concluded that student achievement in the United States lags woefully behind that in many comparable industrialized nations, that this shortcoming threatens the nation's economic future, and that these test results therefore demand radical school reform that includes importing features of schooling in higher-scoring countries.

This report challenges these conclusions. It focuses on the relevance of comparing U.S. *national* student performance with average scores in other countries when U.S. students attend schools in 51 separate education systems run not by the federal government, but by states (plus the District of Columbia). To compare achievement in states with each other and with other countries, we use newly available data for student mathematics and reading performance in U.S. states from the 2011 TIMSS and 2012 PISA, as well as several years of data from the National Assessment of Educational Progress (NAEP).[4] In particular, we use information on mathematics and reading performance of 15-year-olds from the PISA data, information on mathematics performance in 8th grade from the TIMSS data, and information on mathematics and reading performance of students in 4th and 8th grade from the NAEP data.

We conclude that the most important lessons U.S. policymakers can learn about improving education emerge from examining why some U.S. states have made large gains in math and reading and achieve high average test scores. The lessons embedded in how these states increased student achievement in the past two decades are much more relevant to improving student outcomes in other U.S. states than looking to high-scoring countries with social, political, and educational histories that differ markedly from the U.S. experience. No matter how great the differences among U.S. states' social and educational conditions, they are far smaller than the differences between the United States as a whole and, say, Finland, Poland, Korea, or Singapore. As such, this report starts the process of delving into the rich data available on student academic performance in U.S. states over the past 20 years—and shows that the many major state successes should be our main guide for improving U.S. education.

The report is organized around three main arguments:

1. Policymakers are not correct in concluding—based on international tests—that U.S. students are failing to make progress in mathematics and reading.

- Country averages not adjusted for major national differences in students' family academic resources (such as the number of books in the home or mother's education level) mistakenly attribute U.S. students' performance entirely to the quality of U.S. education. After adjusting for these factors, U.S. students perform considerably better than the raw scores indicate.[5]
- Focusing on national progress in average test scores obscures the fact that socioeconomically disadvantaged U.S. students in some states have made very large gains in mathematics on both the PISA and TIMSS—gains larger than those made by similarly disadvantaged students in other countries.
- Student performance in some U.S. states participating in international tests is at least as high as in the high-scoring countries. Additionally, TIMSS gains made by students in several states over the past 12 years are much larger than gains in other countries. Specifically:
    - Students in Massachusetts and Connecticut perform roughly the same on the PISA reading test as students in the top-scoring countries (i.e., Canada, Finland, and Korea)[6] and high-scoring newcomer countries (i.e., Poland and Ireland), and higher than students in the post-industrial countries (i.e., France, Germany, and the United Kingdom). Socioeconomically advantaged students in Massachusetts score at least as well in mathematics as advantaged students in high-scoring European countries.
    - On the 2011 TIMSS, advantaged students in Connecticut, Massachusetts, Minnesota, North Carolina, Indiana, and Colorado performed at least as well in mathematics as their counterparts in high-scoring countries/provinces such as Quebec, England, and Finland.
    - Over 1999–2011, students in Massachusetts and North Carolina made average TIMSS mathematics gains at least as large as students' average gains in Finland, Korea, and England. Over 1995–2011, students in Minnesota made TIMSS mathematics gains similar to those in Korea and larger than those in England.

2. It is extremely difficult to learn how to improve U.S. education from international test comparisons.

- Policy recommendations based on the features of high-scoring countries' education systems, or on the education reforms in countries making large gains on international tests, are overwhelmingly correlational and anecdotal. There is no causal evidence that students in some Asian countries, for example, score higher on international tests mainly because of better schooling rather than large investments made by families on academic activities outside of school.
- Reforms in countries such as Germany and Poland, with big gains in PISA scores but with early vocational tracking, seem to have little applicability to the U.S. system. Such differences between the educational cultures of other countries and of the United States make it difficult to draw education policy lessons from student test scores.

3. Focusing on U.S. states' experiences is more likely to provide usable education policy lessons for American schools than are comparisons with higher-scoring countries (such as Korea and Finland).

- There are vast differences between the conditions and contexts of education in various countries. In contrast, among U.S. states, school systems are very similar to one another, teacher labor markets are not drastically different, and education systems are regulated under the same federal rules. If students with similar family academic resources in

some states make much larger gains than in other states, those larger gains are more likely to be related to specific state policies that could then be applied elsewhere in the United States.

- We analyze relative state performance using data from the NAEP mathematics and reading achievement tests over 2003–2013 in all states and over 1992–2003 in most states. We adjust scores to control for differences in the composition of students in each state (e.g., family characteristics, school poverty, and other factors); while this reduces the variation in scores among states, significant differences remain.
    - In general, state gains in mathematics were larger than in reading. However, there were large variations in gains across states.
        - Over 1992–2013, the average annual increase in NAEP 8th grade adjusted mathematics scores in the top-gaining 10 states was 1.6 points per year—double the 0.8 point annual adjusted gain in the bottom-gaining 10 states.
        - Over 2003–2013, a number of states made significantly large gains in 8th grade mathematics and reading. These include Hawaii, Louisiana, North Carolina, Massachusetts, and New Jersey. States that made large reading gains were not necessarily the same states with large mathematics gains. For example, Texas made large gains in mathematics but not reading.
        - Gains for states with improvements were not necessarily a function of the initial level of student performance. Some of these states started from a low level of student performance, others from a middle level, and others from high levels. For example, students in Hawaii, the District of Columbia, Louisiana, and North Carolina started with low adjusted test scores in 8th grade mathematics and made large gains since the 1990s; students in Massachusetts, Vermont, and Texas started with relatively higher test scores and also made large gains.
- There are undoubtedly different lessons to be learned in states where students made gains from initially low beginnings and from initially high beginnings, or where students made gains on one subject matter but not the other. The next stage for researchers should be to determine what those lessons are. This paper begins this process.
    - When we tested for possible explanations for these variations, we found that students in states with higher child poverty levels perform worse in 8th grade mathematics even when controlling for individual student poverty and the average poverty of students attending a particular school. Thus, students with similar family resources attending school in high-poverty states tend to have lower achievement than students in low-poverty states. We also found that students in states with stronger accountability systems do better on the NAEP math test, even though that test is not directly linked to the tests used to evaluate students within states.
    - As a suggestive strategy for further (qualitative) policy research, we paired off states with different patterns of gains in 8th grade math. This reveals, for example, that 8th grade students in Massachusetts made much larger gains after 2003 than students in Connecticut, that students in New Jersey made larger gains than students in New York after 2003, and that students in Texas already started out scoring higher in 8th grade math in 1992, but still made larger gains over 1992–2013 than students in California, especially after 2003.

- These and other comparison groups could provide important insights into the kinds of policies that enabled students in some states to make much larger adjusted gains in math scores than students in states that are geographically proximate and/or that have similar population sizes.

The tables and figures referenced in the text can be found at the end of the report.

Do U.S. students perform poorly on international tests?

U.S. students score lower, on average, on international tests than students in other developed countries. This has led many policymakers and journalists to conclude that American student achievement lags woefully behind that in many comparable industrialized nations, that this shortcoming threatens the nation's economic future, and that the United States therefore needs radical school reform.

Upon release of the 2011 TIMSS results, for example, U.S. Secretary of Education Arne Duncan called them "unacceptable," saying that they "underscore the urgency of accelerating achievement in secondary school and the need to close large and persistent achievement gaps" (Duncan 2012). A year later, when the 2012 PISA results were released, he stated, "The big picture … is straightforward and stark: It is a picture of educational stagnation." He continued, "The problem is not that our 15-year-olds are performing worse today than before. The problem instead is that they are not making progress. Yet students in many nations … are advancing, instead of standing still" (Duncan 2013). He highlighted the urgent nature of addressing this lack of progress by going on to argue, "In a knowledge-based, global economy, where education is more important than ever before, both to individual success and collective prosperity, our students are basically losing ground. We're running in place, as other high-performing countries start to lap us" (Duncan 2013).

Is Secretary Duncan correct in his conclusions? In 2013, the Economic Policy Institute published a comprehensive report, *What Do International Tests Really Show About American Students' Performance?* (Carnoy and Rothstein 2013), that criticized the common interpretation of U.S. students' performance on international tests. Carnoy and Rothstein (2013, 7) wrote that the interpretation was "…oversimplified, exaggerated, and misleading. It ignores the complexity of the content of test results and may well be leading policymakers to pursue inappropriate and even harmful reforms that change aspects of the U.S. education system that may be working well and neglect aspects that may be working poorly."

The report argued that by focusing on country averages not adjusted for major national differences in students' family academic resources—such as the number of books in the home or mother's education level—policymakers and journalists mistakenly attributed the poor performance of U.S. students entirely to the quality of U.S. education. The results from PISA and TIMSS tests suggest that when adjusted for differences in the family academic resources of samples in various countries, middle school and high school students in the United States *as a whole* do reasonably well in reading compared with students in large European countries, but do not score as highly in mathematics either on the TIMSS or the PISA as students in other developed countries, including larger European countries such as France and Germany (Carnoy and Rothstein 2013). This is particularly true for U.S. middle and highly advantaged students.

The report also argued that focusing on "national progress in average test scores" in one year failed to differentiate gains over time for disadvantaged and advantaged students and how these compared with gains over time for similarly advan-

taged or disadvantaged students in other countries. The report's final argument was that different international and national tests produced different pictures of mathematics achievement gains by U.S. students in the same 1999–2011 period, and that these differences suggested using caution when basing calls for education reform upon the results of any single test. Specifically, when adjusted for the varying family academic resources of the students sampled, U.S. students' average TIMSS mathematics scores increased substantially in the 12 years between 1999 and 2011; the same was not true for adjusted U.S. scores on the PISA mathematics test over 2000–2009. Because the full range of knowledge and skills that we describe as "mathematics" cannot possibly be covered in a single brief test, the report urged policymakers to examine carefully whether an assessment called a "mathematics" test necessarily covers knowledge and skills similar to those covered by other assessments also called "mathematics" tests, and whether performance on these different assessments can reasonably be compared.[7]

The most recent PISA (2012) and TIMSS (2011) results discussed below confirm Carnoy and Rothstein's earlier findings. Compared with students in other countries, U.S. students as a whole continue to perform better, on average, on the PISA reading test than in mathematics, and the results in mathematics suggest that U.S. students, on average, are performing well below students in many other countries. The results also continue to show that when scores on the PISA and TIMSS are adjusted for differences in family academic resources of students in the samples, U.S. students do considerably better than the raw scores indicate, and the results show that disadvantaged U.S. students have made large gains in mathematics on both tests since 1999–2000.

Further, the latest PISA and TIMSS tests included results for a number of U.S. states. Three U.S. states (Massachusetts, Connecticut, and Florida) participated in the PISA in 2012, and nine U.S. states (Alabama, California, Connecticut, Colorado, Florida, Indiana, Massachusetts, Minnesota, and North Carolina) participated in the 2011 TIMSS. The state data give a much more varied picture of U.S. student performance. Students in some of these states performed at very high levels in all subjects, as high as students in the highest-scoring countries. Conversely, students in others performed considerably below the OECD average.

Thus, the PISA and TIMSS results provide a more mixed picture of U.S. student performance than Secretary Duncan claims, with low average scores in mathematics and middle-of-the-road scores in reading for the United States as a whole, but large mathematics gains for disadvantaged students in the past 12–14 years, and a number of U.S. states where students perform at least as well as students in higher-scoring countries, with much larger gains in mathematics in the past decade for students in these states than for students in other countries. Secretary Duncan can claim "stagnation" and "lack of progress," but the data we present below show that the truth is more nuanced.

Comparing U.S. student average performance on the 2012 PISA

To simplify our comparisons of national average PISA scores and of these scores disaggregated by family academic resources (FAR),[8] we focus on comparing students in the United States with students in eight other countries—Canada, Finland, South Korea, France, Germany, the United Kingdom, Poland, and Ireland.

We refer to three of these countries (Canada, Finland, and South Korea) as "top-scoring countries" because they score much better overall than the United States in reading and math—about one-third of a standard deviation better. Canada, Finland, and South Korea are also the three "consistent high-performers" that Secretary Duncan highlighted when he released the U.S. PISA 2009 results (Duncan 2010).

We call three others (France, Germany, and the United Kingdom) "similar post-industrial countries" because they score similarly overall to the United States. They also are countries whose firms are major competitors of U.S. firms in the production of higher-end manufactured goods and services for world markets. Their firms are not the only competitors of U.S. firms, but if the educational preparation of young workers is a factor in national firms' competitiveness, it is worth comparing student performance in these countries with student performance in the United States to see if these countries' education systems—so different from that in the United States—play a role in their firms' success.

We call the remaining two comparison countries (Poland and Ireland) "high-scoring newcomer countries" because they have been cited by the Organization for Economic Cooperation and Development (OECD), the sponsor of PISA, for their high performance in PISA scores, and were middle-income European countries in the 1990s that have had relative economic success since (although Ireland's economy went into a sharp downturn in the 2008 recession).

Table 1 shows that, on average, U.S. performance on the 2012 PISA (498 points in reading and 481 points in math) was substantially worse than performance in the top-scoring countries and high-scoring newcomer countries in both math (49 and 28 points less than in the two groups, respectively) and reading (30 and 23 points less than those groups, respectively), was about the same as performance in the similar post-industrial countries in reading, and was substantially worse than performance in the similar post-industrial countries in math (19 points lower).

How we describe PISA scores in this report

PISA is scored on a scale that covers very wide ranges of ability in math and reading. When scales were created for reading in 2000 and for math in 2003, the mean for all test takers from countries in the OECD was set at 500, with a standard deviation of 100. When statisticians describe score comparisons, they generally talk about differences that are "significant." Yet while "significance" is a useful term for technical discussion, it can be misleading for policy purposes, because a difference can be statistically significant but too small to influence policy. Therefore, in this report, we avoid describing differences in terms of statistical significance. Instead, we use terms like "better (or worse)" and "substantially better (or worse)" (both of which are significantly better for statistical purposes), and "about the same." We also sometimes use "substantially higher (or lower)" interchangeably with "substantially better (or worse)," etc. In general, in this report we use the term "about the same" to describe average PISA score differences that are less than 8 scale points, we use the term "better (or worse)" to describe differences that are at least 8 scale points but less than 18 scale points, and we use the term "substantially (or much) better (or worse)" to describe differences that are 18 scale points or more.

Eighteen scale points in most cases is equivalent to about 0.2 standard deviations. Policy experts generally consider an intervention that is 0.2 standard deviations or more to be an effective intervention; such an intervention, for example, would improve performance such that the typical participant would now perform better than about 57 percent of all participants performed prior to the intervention.

And in the case of trends, we sometimes speak of scores that were "mostly unchanged," a phrase with identical meaning as "about the same." Further, it is difficult to interpret and compare the results from various assessments because each test has its own unique (and arbitrary) scale.

However, the variation among U.S. states is large. In reading, students in Massachusetts (527 points) and Connecticut (521 points) perform about the same as students in the top-scoring and high-scoring newcomer countries and higher than students in the post-industrial countries. (Indeed, in reading, Massachusetts students only perform worse than students in Korea.) However, students in Florida (492 points) perform worse or substantially worse than all the comparison groups of countries. In mathematics, where the U.S. average is substantially lower than in all three comparison groups of countries, students in Massachusetts and Connecticut (514 and 506 points, respectively) perform as well as or better than students in all three post-industrial countries and in one of the high-scoring newcomers, Ireland. They only perform worse than the average of the group of high-scoring countries and Poland. On the other hand, students in Florida (467 points) perform substantially worse in math (as was the case in reading) than students in all three comparison groups of countries.

Comparing the 2012 PISA performance of U.S. students with different levels of family academic resources

We next disaggregate scores in the U.S. states, in the United States as a whole, and in the eight comparison countries by an estimate of the family academic resource status of test takers, dividing them into six groups, from the least to the most advantaged.[9] We refer to these as Group 1 (lowest FAR), 2 (lower FAR), 3 (lower-middle FAR), 4 (upper-middle FAR), 5 (higher FAR), and 6 (highest FAR). We also refer to Groups 1 and 2 together as disadvantaged students, to Groups 3 and 4 together as middle-class students, and to Groups 5 and 6 together as advantaged students.

From **Tables 2A and 2B** we can observe that more U.S. 15-year-olds (40.2 percent) and Florida 15-year-olds (48.3 percent) are in the disadvantaged FAR groups (Groups 1 and 2) than in any of the eight comparison countries or Massachusetts and Connecticut. We can therefore see why comparisons that do not control for differences in family academic resource distributions between countries and states may differ greatly from those that do. Yet, even Connecticut and Massachusetts have a higher percentage of 15-year-olds in the disadvantaged groups (29.1 and 32.0 percent) than the high-scoring countries—Canada, Finland, and Korea (25.2, 22.6, and 12.9 percent, respectively). The proportion of advantaged students in the Massachusetts and Connecticut samples (21.7 and 24.2 percent, respectively) are more similar to the proportions in all the comparison countries except Poland, which has a somewhat lower proportion (17.8 percent).

Any comparison of average performance in a country as a whole that is not adjusted for these differences is hardly a meaningful measure of learning differences possibly due to quality of schooling. Also, in comparing across countries, differences in the student body composition in the different countries need to be accounted for. We know that family-resource-disadvantaged students universally have lower academic performance than more advantaged students. Tables 2A and 2B allow us to compare the scores of students by FAR group in the United States as a whole (and in the three U.S. states) with students in the same FAR group in other countries. The tables also show the typical "unadjusted" PISA score for each state and country and the score adjusted for differences in the proportion of students in each FAR group

in each country's sample (these scores are depicted graphically in **Figure A**). We make the adjustment by assuming that each country and state has the same proportion of students in each FAR group as the U.S. sample. This "adjusted" score offers a more accurate comparison of how students' schooling influences academic performance.

These results confirm earlier findings suggesting that disadvantaged FAR students in the United States compare more favorably to their foreign counterparts than do U.S. middle and highly advantaged students.[10] The results also show that although adjusting average PISA mathematics and reading scores reduces the difference between U.S. and comparison country averages, scores remain lower in the United States than in the high-scoring and newly high-scoring countries.

However, students across all FAR groups in higher-scoring U.S. states, such as Massachusetts and Connecticut, do at least as well in reading as students in four of the highest-scoring countries (Canada, Finland, Poland, and Ireland) on the 2012 PISA test (see Table 2A). It is important to note that in sharp contrast to the United States as a whole and in sharp contrast to lower-scoring (and poorer) states such as Florida, Massachusetts and Connecticut's substantial proportion of advantaged students (more than 20 percent of the sample in each state) generally do at least as well in mathematics as advantaged students in the highest-scoring countries (including the newly high-performing), and in the post-industrial countries.[11] [12] The only exception is that the most advantaged students in Korea (i.e., those in Group 6) score substantially higher in math than the most advantaged students in Connecticut and Massachusetts.

Even so, Massachusetts students do slightly better than Korean students on the PISA reading test when we adjust for differences in student FAR (the difference is 7 points, which we typically classify as "about the same") (Table 2A). In addition, there is one group in Massachusetts that may be comparable to students in Korea in terms of out-of-school activities: self-identified Asian-origin students. Students who identified themselves as being of Asian origin in the Massachusetts PISA sample scored 569 on the 2012 PISA mathematics test, significantly higher than the unadjusted average in Korea (554), and about the same as students in Singapore (573).

Comparing changes in the PISA performance of low FAR and high FAR students in the United States with their counterparts in other countries, 2000–2012

The results in **Table 2C** show that although average U.S. PISA scores stagnated over 2000–2012, U.S. disadvantaged students (those in FAR Groups 1 and 2) made very large gains compared with similarly disadvantaged students in a number of post-industrial and high-scoring countries (in line with the findings in Carnoy and Rothstein 2013). A positive number in the table indicates that U.S. students in a particular FAR group gained that many scale points in reading or math over 2000–2012 compared with the same FAR group in the comparison country. A negative number means that U.S. students fell further behind their counterpart FAR group in the comparison country. Thus, compared with Group 1 and 2 students in other countries, U.S. students made relative gains on the PISA mathematics test of between 13 scale points (compared with Group 2 in Ireland) to 60 scale points (compared with Group 1 in Finland) between 2000 and 2012, and made even larger relative gains on the reading test. However, U.S. Group 5 students lost ground to students in every country except the U.K. in both reading and math.[13] Furthermore, U.S. students in every group lost considerable ground to their corresponding group students in Germany and Poland.

The results in Table 2C also confirm that high FAR students in the United States have been performing steadily worse on the PISA compared with their counterparts in other countries (a conclusion of Carnoy and Rothstein 2013). This

has important implications for U.S. policymakers—implications that require a more nuanced analysis of where to focus in trying to raise U.S. students' mathematics performance.[14]

Comparing average 2011 TIMSS performance of students in U.S. states, the United States as a whole, and other countries

Table 3A compares 2011 TIMSS mathematics results for the United States as a whole, the participating Canadian provinces (Alberta, Ontario, and Quebec; Canada as a whole did not participate), Finland, South Korea, and England. **Table 3B** presents results for the nine U.S. states that participated individually in the 2011 TIMSS (Massachusetts, Minnesota, Connecticut, Indiana, Alabama, Colorado, North Carolina, California, and Florida).

Column 6 of Tables 3A and 3B displays the published average 2011 TIMSS scores of each country, state, or province. Column 7 of both tables reweights the average scores, assuming that each country, state, or province had a FAR distribution similar to that of the United States nationwide. It shows that adjusting for FAR composition makes little difference in the overall average scores of most countries, provinces, and states. The greatest differences are in the cases of Ontario, Korea, Massachusetts, Minnesota, and Alabama.

The results in Tables 3A and 3B add support to the results we found in the PISA comparisons, contradicting the general lament that the United States performs relatively poorly in mathematics. After adjusting for FAR distribution differences by applying the overall U.S. FAR distribution to the various countries, provinces, and states, students in seven of the nine U.S. states that took the TIMSS in 2011 performed better than students in all but Korea and Quebec. Students in Korea perform substantially better in every FAR group than in any U.S. state, including Massachusetts.

Since Finland has been held up as having an exemplary education system, it is noteworthy that each FAR group in Massachusetts, Minnesota, and North Carolina outperformed or substantially outperformed comparable students in Finland. Students in Indiana, Colorado, and Florida performed the same as or better than students in Finland in every FAR group.

Overall, students in Massachusetts in every FAR group performed substantially better than comparable students in the three Canadian provinces, Finland, and England. The exceptions are the Group 1 students in Quebec, who performed as well as comparable students in Massachusetts. Students in North Carolina and Minnesota also performed as well as or better than comparable students in these provinces and countries in all but the lowest (North Carolina, Minnesota) and lower (Minnesota) FAR groups.

As in the case of Massachusetts and Connecticut in the 2012 PISA results, on the 2011 TIMSS mathematics test higher FAR students (Groups 4 and 5/6) in states such as Massachusetts, Minnesota, and Connecticut outperformed their counterparts in comparison provinces and countries to a greater degree than did disadvantaged students (Groups 1 and 2) in those states. As in the 2012 PISA, the one exception to this trend is Korea. Excluding the Korean exception, this overall result suggests two possibilities. The first is that advantaged students in higher-scoring U.S. states are being more adequately prepared in mathematics, relative to disadvantaged students, when compared with other high-scoring countries and provinces. The second is that there are factors in the disadvantaged student groups in many states (including Massachusetts, Minnesota, and North Carolina) other than the number of books in the home that may negatively affect

test scores. These factors—such as English language proficiency and some factors mediated by race—may be less prevalent in Canada's provinces, Finland, or England.

These TIMSS results for several other states, in addition to Connecticut and Massachusetts, that also scored highly on the PISA test—Colorado, Indiana, Minnesota, and North Carolina—suggest that students in various parts of the United States, particularly advantaged students, are performing at least as well in mathematics as their counterparts in "high-scoring" countries such as Canada and Finland, but not as well as advantaged students in Korea.

Changes in U.S. states' TIMSS scores over time

Three U.S. states took the PISA test for the first time in 2012, but a number of U.S. states have participated in the TIMSS test since 1995—Connecticut, Massachusetts, Minnesota, North Carolina, Indiana, Missouri, and Oregon. This allows us to observe how well students in various states performed on the TIMSS mathematics test over 1995–2011, and we can compare these changes to those in several of our comparison countries/provinces that also took the TIMSS over this period.

Table 4 shows the results for those states that participated in more than one administration of the TIMSS. Students in Missouri and Oregon, which participated only in 1995 and 1999, scored lower in 1999 than in 1995. Five states made gains: Minnesota (over 1995–2011) and Connecticut, Indiana, Massachusetts, and North Carolina (over 1999–2011). The gains in North Carolina and Massachusetts are particularly large.[15]

Table 5A compares changes in TIMSS scores across FAR groups in 1999–2011 in Connecticut, Massachusetts, Indiana, and North Carolina with changes in the same FAR groups in Finland, Korea, England, and the United States as a whole over this period. **Table 5B** compares changes in Minnesota over 1995–2011 with changes in Korea, England, and the United States as a whole over this period.

Table 5A shows that the increases in TIMSS mathematics test scores in all FAR groups in Massachusetts and North Carolina were very large over 1999–2011, about 50 scale points, or one-half a standard deviation. Mathematics gains in Connecticut and Indiana were smaller, ranging between 0.1 and 0.3 standard deviations in Connecticut and about 0.2 to 0.4 standard deviations in Indiana. In the United States as a whole, advantaged students (Groups 5/6) made smaller gains than disadvantaged and middle FAR students. However, in these four states, this pattern of gains was shared only in Indiana, and even in Indiana, the gain for advantaged students was almost 0.3 of a standard deviation. In this same period, students' performance in Finland stayed about the same, but the score of Group 1 fell 18 points, or about 0.2 standard deviations. Students in Korea also made gains, but these were much smaller than those in Massachusetts and North Carolina, and more similar to those in Connecticut and Indiana. The pattern of gains in Korea tended to be larger for advantaged students than disadvantaged students.

As shown in Table 5B, students in Minnesota made large gains in 1995–2011 across all groups, with somewhat larger gains in middle and disadvantaged groups. The gains were about the same as in the United States as a whole, Korea, and England.

The distribution of students among FAR groups in the U.S. TIMSS samples changed radically over 1995–2011. It was more similar to the distributions in Korea's and England's samples in 1995 and more similar to Finland's, Korea's, and

England's samples in 1999 than in 2011. The percentage of disadvantaged students in the U.S. sample increased considerably in the years after 1995 and 1999.

In column 7 of Tables 5A and 5B (shown graphically in **Figures B** and **C**), we estimate the hypothetical score each state and country would have had were the distribution of the sample among FAR groups the same as the distribution in the United States as a whole in 2011. This adjustment increases the gain for every state, for the United States as a whole, and for England. Finland's loss stays about the same, and Korea's gain from 1999 to 2011 drops significantly (however, Korea's gain from 1995 to 2011, shown in Table 5B, stays about the same).

Such very large gains in mathematics for students in U.S. states, particularly those in Massachusetts and North Carolina, do not conform to the characterization of U.S. education as failing. This more positive image of U.S. education in several of these states is reinforced by the fact that, as we showed in Tables 3A and 3B, students in several states (Massachusetts, Minnesota, and North Carolina)—once their average TIMSS scores and those of comparison countries are adjusted to a common FAR sample distribution—perform at least as well in mathematics as students in the Canadian provinces, Finland, and England (but worse than in Korea).

Why it is difficult to learn much about improving U.S. schools from international test comparisons

Over the past several years, the controversy around the validity and politics of using international tests to draw conclusions for educational policy and practice has intensified. The critiques cover a broad range of issues. Some question the reliability of the PISA test and the PISA rankings (Stewart 2013). Others cast doubt on the representativeness of a key PISA sample, Shanghai, continuously held up by the OECD as a symbol of educational excellence (Loveless 2013, 2014; Harvey 2015).[16] The discussions about the validity of international tests as measures of students' knowledge and the representativeness of PISA samples reveal an important aspect of these tests.

The most cogent of the recent critiques is that international agencies—especially the OECD—are often too quick to use international test score data to argue that particular education policies are the reason test scores are high in some countries and low in others. A main policy recommendation coming out of international comparisons is to copy or adapt the policies of higher-scoring countries (see OECD 2011). For example, because students in some East Asian countries and cities—such as Korea, Japan, Singapore, and, most recently, Shanghai—achieve such high test scores, the OECD and others consistently feature these as exemplary education systems. Some reasons given for educational quality in East Asia are the high level of teacher skills, high teacher pay, and, in some countries, such as Korea, a rather equal distribution of students from different social class backgrounds across schools.[17] Yet these explanations are only backed by correlational, not causal, evidence.

Furthermore, out of 55 countries that have taken the PISA mathematics test over a number of years, only 18 trend upward. Of these 18, about half are low-scoring developing countries with levels of educational and economic development very different from those of the United States. The OECD has focused heavily on the high-scoring countries and big gainers, but it has failed to balance the discussion with explanations for why students in countries with "good" school systems—such as Finland, Australia, New Zealand, Canada, Belgium, Hungary, the Czech Republic, and Sweden—did significantly worse on the PISA mathematics test in 2012 than in 2003 (OECD 2013a, Figure I.2.16).

U.S. policymakers seem to be a particular target for recommendations drawn from the PISA data on how to improve U.S. education, and these recommendations further illustrate the pitfalls of using other countries to draw lessons for U.S. policy. In the wake of the PISA 2009 score release, Secretary Duncan requested that the OECD prepare a report on lessons for the United States from international test data. In that report, *Lessons from PISA for the United States*, the OECD advised Duncan to follow the lead of educators in Ontario, Shanghai, Hong Kong, Japan, Germany, Finland, Singapore, and Brazil, in each case drawing "lessons" of how schooling in those countries/provinces/cities brought students to high or higher levels of performance (OECD 2011). In 2013, OECD produced a second *Lessons from PISA* aimed at U.S. policymakers, this time making more general recommendations drawn from correlational analyses of the 2012 PISA data (OECD 2013d).

As we have shown, there is reason to agree with international testing proponents that the U.S. education system could be improved to teach students mathematics better, and, less urgently, to make U.S. students into better readers. Detailed studies about why other countries' education systems do well or badly can certainly provide many points of discussion for U.S. educators and policymakers about how school systems are organized, about curriculum differences, and about teacher training, among many other themes. For example, 20 years ago, based on international comparisons of 1995 TIMSS test scores, studies were able to argue convincingly that students in the United States did poorly on the 8th grade TIMSS mathematics test mainly because only one-quarter of U.S. students were exposed to algebra and an even lower share were exposed to geometry by the 8th grade (Schmidt et al. 2001). Scholars also argued that the U.S. math curriculum was a "mile wide and an inch deep" (Schmidt et al. 1997). This was a wise use of international comparisons to make policy changes in mathematics. Likewise, there may also be useful research in other countries on the impact of education policies on student outcomes. However, as we shall see, looking to other countries to find solutions for education issues involving nationally or even regionally specific and complex student-teacher-administrative interactions is certainly not the only option that U.S. education policymakers can contemplate.

Why some of the policy recommendations are not feasible

Education systems develop in social and political contexts and are inseparable from those contexts. Families in some cultures are more likely to put great emphasis on academic achievement, particularly on achievement as measured by tests. They act on this emphasis by investing heavily in their children's out-of-school academic activities, including "cram courses," which focus on test preparation (Ripley 2012; Bray 2006; Wantanabe 2013; Bray and Lykins 2012; Byun 2014).[18] In a world that puts high value on test scores, perhaps such intensive focus on children's academic achievement should be applauded. However, whether it is a good choice for middle and high schoolers to spend most of their waking hours studying how to do math and science problems, and whether it is likely that families in other societies would buy into this choice for their children, are highly controversial questions and certainly only somewhat related to the quality of schooling received by students in a particular society.[19]

There is some evidence that mathematics instruction in certain Asian countries is better than in the United States. Comparisons of mathematics instruction in Japanese and U.S. schools (Hiebert et al. 1999) support the "mile wide, inch deep" critique of U.S. mathematics education made almost 20 years ago (Schmidt et al. 1997). There is also evidence that greater "exposure to formal mathematics" has a positive effect on TIMSS and PISA mathematics performance (Schmidt et al. 2001; Schmidt et al. 2015; Carnoy et al. 2015a). However, the relevant questions are whether there is something specific we can learn from Japanese school practices that we know will increase U.S. student mathe-

matics performance, and whether greater exposure to formal mathematics is taking place mainly in schools.[20] As noted, if greater exposure outside of school is the more important reason that students in the PISA sample, in, say, Korea, Japan, Singapore, Hong Kong, or Singapore do better on the PISA mathematics test than students in the United States, we would need to change core behavior patterns in most U.S. families. Simply improving mathematics instruction in school would not help much if spending significantly more time on mathematics outside school is the main reason 15-year-olds in Korea score higher on the PISA test.

Finland has also been touted as having a model education system, mainly because of its highly trained teachers and the autonomy teachers and principals in the country allegedly have in their classrooms and schools.[21] Teacher education and classroom teaching in Finland indeed seem very good, but neither the OECD nor the Finns ever offered any evidence approaching causality to support these claims.[22] Furthermore, performance on PISA reading and mathematics tests by students in the country's lower FAR groups has deteriorated considerably over the past decade (Table 2C), though this decline has not been addressed by PISA analyses.

Two recent examples of "successful" reforms featured in OECD reports are those undertaken by Germany and Poland (OECD 2013b, Volume II, Box II.3.2; OECD 2013c, Volume IV, Box IV.2.1). Students in those two countries scored near the OECD average in 2003 and have made large gains since. One study of the Polish reforms argues that Poland's 1999 reform postponing vocational tracking from the 9th to the 10th grade lifted by one standard deviation the PISA reading scores of students who would have gone to vocational school in the 9th grade. The study argues that this explains much of Polish reading test score gains in 2000–2006 (Jakubowski et al. 2010). Yet, thanks to a special follow-up sample in Poland, that same study is also able to show that in 10th grade, the 9th grade cohort entering the vocational education track "lost" the gains they had made in 9th grade.[23]

Germany had a PISA-reported increase in its sample of disadvantaged students in the 2000s, making the achievement in that decade of substantial gains for such students more interesting and impressive. Nevertheless, the gain in average test scores from 2000 to 2009 apparently came from gains made by children from Slavic-country immigrant families. The gains of ethnic German students were negligible (Stanat et al. 2010). The reported concentration of German student performance gains in first- and second-generation immigrants, most from Russia and Eastern Europe, raises questions about whether school reforms were related to such gains or whether lessons learned in Germany from educating Russian and Eastern European immigrants are applicable to the U.S. context, where most low-scoring immigrants are from Mexico and come to the country with considerably less family academic resources. An OECD report on lessons for the United States from other countries discusses German reforms but concedes that there seems to be no empirical link between those reforms and German test score gains (OECD 2011).[24]

These are just a few cases illustrating how a lack of relevance and causal evidence does not impede the OECD and others from drawing conclusions from PISA data on what works to increase student test scores. The OECD has made a regular practice of recommending what countries and even schools should do to increase their students' learning even though there is no link to particular country situations and no causal evidence for these claims (OECD 2013a, Chapter 6).

When evaluating such recommendations, U.S. policymakers need first to ask whether they are relevant to the U.S. context—whether the experiences described above, such as those in Asia, Finland, Germany, and Poland, are meaningful

for U.S. educational conditions. Secondly, U.S. policymakers need to ask whether the analysis underlying the recommended intervention or reform makes a reasonable inferential case that the intervention or reform is linked to improved education outcomes. It is a challenge to find any OECD recommendations in the two *Lessons from PISA for the United States* volumes that meet these criteria. For all these reasons, it is challenging to learn about improving U.S. schools from comparisons based on international tests.

Should U.S. policymakers look outward or inward?

Because of all these issues of comparability, the PISA and TIMSS results in various U.S. states are particularly interesting. Our analysis in the previous section suggests that the average of even usually lower-scoring advantaged students in the United States *as a whole* may be a poor measure of success or failure of such students in *many of the administrative units that actually administer U.S. education*. In a number of states that have participated in international tests, advantaged students do at least as well in mathematics—in which students in the United States as a whole do not perform very well—as advantaged students in high-scoring PISA and TIMSS countries. At the same time, advantaged students in many other states do worse than their counterparts in large industrial European countries and much worse than students in high-achieving countries.

Thus, the question U.S. policymakers and educators should ask is whether they need to turn to the rest of the world to find out what will make relevant contributions to improving U.S. education—or whether they should look within the United States' varied multi-educational state-based systems to find these answers.

The case for looking inward across states

The case for looking inward, across U.S. states, is compelling. When considered carefully, the concept of a "United States education system" is largely a construct of agencies administering international tests and the U.S. Department of Education. Education policies are, in the environment of the past 20 years, largely formulated by states and local school districts, even though federal legislation and programs such as the Civil Rights Act, Title I, Title IX, No Child Left Behind, Race to the Top, and Common Core are important in holding states and districts to certain requirements and steering them in certain directions. Ultimately, however, implementation varies, and the result is considerable variation in the quality of education systems even among states. As would be expected, student performance in mathematics on international tests varies greatly among U.S. states.

Secondly, successful states' results and experiences are more relevant for drawing policy lessons for state policymakers because the conditions and context of education are more similar among U.S. states than among the United States and other countries. School systems are very similar in U.S. states, teacher labor markets are not drastically different, and the state systems are regulated under the same federal rules. If students with similar family academic resources in some states make much larger gains than in other states, those larger gains are reasonably likely to be related to specific state policies that are *transferable across states*.

If we are to learn lessons for improving 15-year-olds' mathematics and reading performance, we argue that looking at the great number of successful education systems within the United States is far more useful and feasible than turning for lessons to Finland or Germany (which is also not a single education system) or Poland. If test score levels were not very high in the best-performing U.S. states, one could argue against this approach. But this is not the case. Students in

Massachusetts with similar family academic resources do at least as well as students in all but a limited group of Asian countries, and were students in Texas or North Carolina or Vermont to take the PISA test, so probably would they.

Thirdly, as shown, many states have made large gains—particularly in mathematics—since the 1990s. As evidenced by the TIMSS test, in some of these states, these gains have been larger than in countries such as Finland, held up as examples of good education policy. In Indiana, for example, the largest gains on the TIMSS mathematics test appear to be among disadvantaged students, and in others, such as Connecticut, the gains appear to be largest for advantaged students. In three states—Minnesota, Massachusetts, and North Carolina—large gains are spread across all family academic resource groups. Differences in states' policy changes may reveal important lessons for policymakers.

For all these reasons, it makes greater sense to use student performance across U.S. states to understand why students in some are able to achieve high mathematics scores on several kinds of tests than to look to other countries for educational policy direction.

What we can learn from student performance in U.S. states over time

Having made the case that systematically comparing student performance in U.S. states could be highly valuable for education policymakers, we now turn to an analysis of the differences in student achievement in mathematics and reading across U.S. states over the past 10 and 20 years, using microdata available from the National Assessment of Educational Progress (NAEP).[25] We have established that there is considerable variation in student performance among U.S. states on both the PISA and TIMSS tests, and that student gains on the TIMSS mathematics test for five states also vary considerably from the 1990s to 2011. NAEP data allow us to extend this analysis to all U.S. states and to begin to suggest ways to draw lessons from the results to improve student achievement.

Comparing states' performance on the NAEP over 1992–2013 using regression analysis

Among its multiple uses (see, for example, Lubienski 2002; Lubienski and Lubienski 2014), the NAEP, the United States' national assessment test, is the main data source for interstate comparisons. It provides state-level data since 1990 and sufficiently consistent samples of states since 1992. The NAEP is applied to students in specific grades (4th, 8th, and 12th). We focus on the 8th grade results in mathematics and reading but also estimate results for students in the 4th grade to check on the possibility that changes in 4th grade scores in a state drive changes in 8th grade scores four years later. These are called "cohort effects."

Student performance on the NAEP varies considerably across states in each year the NAEP is given. For example, in 8th grade mathematics in 2013, students in the lowest-performing jurisdictions, the District of Columbia and Alabama, averaged mathematics scale scores of 265 and 269. At the top of the spectrum, students in New Jersey scored 296, and students in Massachusetts scored 301. This represents a spread of about one standard deviation in average scores among individual students taking the test and about 0.8 of a standard deviation among state averages.

Just as in our analysis of TIMSS and PISA results across countries and states, average NAEP scores can also differ among states for reasons other than the quality of schooling. One of these is demographics. In states with a higher percentage of students from low-academic-resource families, average scores would tend to be lower. In states with a higher percentage of minorities such as African Americans or Hispanics—groups that because of a history of discrimination or, in the case of recent immigrants, because of limited English language ability, traditionally do less well on such tests—average scores

are likely to be lower. The portion of the lower scores resulting from differences in the composition of students' characteristics (including gender, race/ethnicity, and age), or differences in the composition of students with different levels of family resources (including language spoken in the home, mother's education, and individual poverty, measured by eligibility for free or reduced lunch), cannot reasonably be attributed to the performance of the education system.

Using the NAEP individual- and school-level microdata, we estimate student performance on the NAEP as a function of students' characteristics, students' family academic resources, and variables identifying each state. The coefficients for each state show how students in that state score once we account for that part of the variation in test scores associated with these student and family resource differences among states. We know that part of the variation in observed test scores among students is related to the family-academic-related skills that students bring to school, and that the level and distribution of students' family resources vary across states. Thus, to get a better understanding of how educational quality differences vary across states, we want to take out student demographic variation, regardless of what the ultimate source of those differences across states might be. We call this our Model 1 adjustment.

In addition to adjusting student achievement for student sample demographics, as we were able to do with the international test data, NAEP offers the opportunity to dig further into the sources of state differences. With NAEP, in addition to controlling for student family characteristics, we can control for some school- and teacher-level factors that may differ across states and are likely to be correlated with student academic performance. Specifically, we can control for the concentration in schools of free and reduced-price lunch students and of black or Hispanic students—two types of school-level "peer effects." We call this our Model 2 adjustment. If we want to adjust further to get at how well states are doing given educational resources available in the states, we can also control for the several characteristics of teachers measured by the NAEP in the schools attended by the sampled students, and for whether the school they attended is public or private. We call this our Model 3 adjustment.

Furthermore, unlike PISA or TIMSS data, the NAEP data allow us to track changes in test scores over time in *all* the states (after 2003).[26] If some states are making larger gains in test scores compared with other states and these gains are not the result of favorable changes in student demographics, we could posit that *something* about those states' school systems is contributing to greater student performance gains. If we also adjust the test scores for the several available measures of teacher resources in each state, the coefficients for each state represent a "state effect" on student performance net of student family academic resources, peer effects from the way state education systems concentrate students in schools, and some measurable teacher resources.

Adjusted score rankings for states, 2003–2013

Because all states were required to take all the tests beginning in 2003, estimates for the adjusted scores for each state in both NAEP reading and mathematics are available for 2003–2013. For this period, we compared the relative positions of each state and their relative gains during that period, and we proceed as mentioned above: We first estimated the regressions using the NAEP microdata files, including, stepwise, student characteristics (Model 1), then adding the social class and race concentration of students in schools (Model 2), then, in a third step, adding type of school (private, charter, or public), teacher experience, and whether teachers had majored in the tested subject (mathematics or reading) in undergraduate or graduate studies (Model 3). In order to enter all the states in the regressions as discrete variables, we needed to omit one state as a reference variable. In our estimates we left out California, a large and relatively low-scoring state. The "adjusted state fixed effects" in each of these three regressions were estimated from the coefficient of the states

added as variables in the regressions.[27] This process (using Model 3) is followed to estimate the "adjusted state fixed effects" for students in 4th and 8th grade, in math and reading.[28] [29]

The results for the average state scores are shown in **Tables 6A1** and **6A2** and **Tables 6B1** and **6B2** (the tables show average state scores "corrected" or adjusted for all of the variables included in models 1, 2, and 3 that are relevant to student performance). Table 6A1 shows the rankings of the states by their adjusted mathematics scores for the years 2003, 2007, 2011, and 2013 for 8th grade mathematics, and Table 6A2 displays the same information for 4th grade mathematics. Table 6B1 shows similar results for 8th grade reading, and Table 6B2 displays data for 4th grade reading.

Our adjustments to reported state average NAEP test scores reduce the variation in students' average performance among states, as would be expected. Table 6A1 shows that in 8th grade mathematics, the state with the lowest average score in 2013 was Alabama (278 points), and the highest-scoring state was Massachusetts (298 points). The 20-point difference in adjusted scores was well below the 36-point spread in observed (i.e., unadjusted) scores between D.C. (265) and Massachusetts (301).[30] We obtain similar decreases in the spread of 4th grade math scores. In 2013, there was a 24-point spread between lowest-scoring D.C. (229) and highest-scoring Minnesota (253) in the average observed score, compared with a 14-point spread in adjusted scores between lowest-scoring California (235) and highest-scoring Indiana (249) (the adjusted scores are shown in Table 6A2).

In 8th grade reading in 2013, the spread fell from 29 points in observed scores—lower than in math—between D.C. (248) and Massachusetts (277), compared with a 14-point spread in adjusted scores between lowest-scoring West Virginia (260) and highest-scoring Massachusetts (274) (the adjusted scores are shown in Table 6B1). In 4th grade reading, the spread falls from 26 points between lowest-scoring D.C. (206) and highest-scoring Massachusetts (232) to 22 points in the adjusted scores, between lowest-scoring Hawaii (209) and highest-scoring Florida (231) (the adjusted scores are shown in Table 6B2). If we consider that the high Florida score may be an anomaly, omitting it yields a spread between Hawaii and Maryland or Massachusetts of 18 points—even this lower figure is a high percentage of the observed range. When we adjust only for average family resources of individual students and within schools as a whole, the adjusted range is the same.

The most evident takeaway from the rankings in each year and from the changes in the adjusted rankings over time is that some states start out low in 2003 and remain low in 2013 because their students made relatively small gains in math or reading over the decade. In other states, students made large gains. The gains were not restricted to states that started out with low scores in 2003; improvements were observed among a wide variety of states, and were not uniquely determined by initial scores. These trends are explained below in more detail.[31]

Mathematics gains in states with low, middle, and high initial (2003) levels

In 8th grade mathematics, students in Alabama, California, Michigan, Utah, and West Virginia scored relatively low (275 or below) in 2003 and remained low in 2013 (281 or below) because they made relatively small gains (7–9 scale points). In 4th grade mathematics, students in Alabama, California, Idaho, and Utah also had relatively low scores in 2003 (225–231 scale points) and scored low in 2013 (235–237 scale points). Michigan students were nearer the middle of the distribution (233 points) in 2003 but only made a 3-point gain in 10 years. However, students in West Virginia, Kentucky, and Tennessee began low in 2003 (229–231 points) but made a rather large gain to 241 points, climbing to the middle of the pack in 4th grade math by 2013.

A number of states where students had relatively low scores on the 8th and 4th grade NAEP mathematics test in 2003 made large adjusted gains by 2013. These include Hawaii, Rhode Island, Nevada, Tennessee, Arkansas, Mississippi, and the District of Columbia. Although there are several caveats to this generalization when we compare 4th and 8th grade initial scores, these six states and D.C. can be considered examples of lower-scoring states whose students made large math gains in 2003–2013.

The boundary between a middle- and high-scoring state is somewhat arbitrary, but we chose 277 to less than 280 as the mid-range for the 2003 8th grade math test, and 232 to less than 235 for the 2003 4th grade test. Students in states such as Nebraska, North Dakota, Wyoming, Idaho, Arizona, New York, and Connecticut had mid-level adjusted scores in 2003 on the 8th grade test, and made small gains in 2003–2013. Students in some of these states, such as Arizona, made larger gains on the 4th grade math test, but others, such as the 8th and 4th grade students in Connecticut and New York, hardly made any gains. Michigan 4th graders also were mid-level scorers in 2003 who made little or no gain in those 10 years.

Students in Louisiana, Maryland, Maine, Massachusetts, New Jersey, and Pennsylvania scored at or extremely close to the middle level in the 2003 8th grade test and made large gains (greater than 14 points) over 2003–2013. Students in New Hampshire also made somewhat smaller but substantial gains from a mid-level 2003 starting score. Students in all these states but Louisiana (238) and Massachusetts (236) began at a middle level in 2003 on the 4th grade math test, and all but Louisiana and Pennsylvania students made large gains over 2003–2013.

Of those states whose students were already achieving at a relatively high adjusted level in mathematics in 2003, a number made little or no progress over 2003–2013. These include South Carolina and South Dakota in 8th grade math, and Louisiana in 4th grade math.

Yet, many more states whose students began at a high level on the 2003 8th grade math test made large gains over 2003–2013. These include Delaware, Kansas, Texas, North Carolina, Indiana, and Vermont. Students in all these states made large gains on the 4th grade math test.

Reading gains in states with low, middle, and high initial (2003) levels

Tables 6B1 and 6B2 show 8th and 4th grade adjusted reading scores, respectively, by state in 2003, 2007, 2011, and 2013. In 8th grade reading, Alabama and West Virginia students were low scoring in 2003 and made small (or slightly negative) gains over 2003–2013. In 4th grade reading, Arizona and Utah start out relatively low and stay relatively low—their 4th grade students made no or very small gains over the decade. Students in Hawaii scored very low in 2003 in both 8th and 4th grade reading and by 2013 made an 11-point gain in 8th grade reading and a 7-point gain in 4th grade reading. Yet Hawaii remained among the very lowest-scoring states in both grades from 2003 to 2013.

Eighth grade students in D.C., California, Maryland, Nevada, and Utah started relatively low in 2003 and made substantial gains by 2013 (7–14 scale points). In all but Utah, the 4th graders in these states also made substantial increases. One caveat is that Maryland 4th graders initially scored at the middle level (216 points in 2003); even so, Maryland students made substantial increases in average 4th grade reading scores over 2003–2013.

Students in Arizona and North Dakota had mid-level 8th grade reading scores in 2003 and made small gains over 2003–2013. North Dakota students also made small gains on the 4th grade reading test. However, in a number of other

states besides North Dakota, students scored at the middle level in 2003 and made very low or even negative gains on the 4th grade reading test over 2003–2013. These include Iowa, Nebraska, and New Mexico.

Students in a larger number of states with mid-level scores in 2003 showed substantial increases in their average 8th grade reading scores by 2013. Idaho, Louisiana, Pennsylvania, and Washington all made 6- to 9-point increases in average reading scores. However, of these states, only Pennsylvania showed a substantial (8-point) average increase on the 4th grade test.

A number of states—Mississippi, New York, South Dakota, and Virginia—with high initial average 8th grade scores showed small increases in these average scores over 2003–2013, and all but Virginia showed small or even negative changes in their average 4th grade scores over this period.

Students in a larger number of states with initially high average adjusted scores on the 2003 8th grade reading test made substantial increases by 2013. These include Connecticut, Kentucky, Massachusetts, Florida, New Jersey, and Vermont. But of these, only Massachusetts, Florida, and New Jersey made relatively large increases in their average 4th grade reading scores.

In summary, there are a group of states with relatively small increases in average adjusted NAEP 8th and 4th grade mathematics scores over 2003–2013, and a group of states with relatively large increases. Similarly, there are a group of states with relatively small increases in average adjusted 8th and 4th grade reading scores, and a group of states with relatively large increases. The overlap between states that made small increases in all four tests is considerable, but some states do better in math than reading, and vice-versa. Texas is one example of a state where students have scored very high in math and made large gains, yet, while still scoring reasonably high in reading, have made much smaller gains.

State rankings by gain in 8th grade adjusted mathematics score over 1992–2013

Table 7 shows how NAEP 8th grade mathematics test scores by state—adjusted just for student characteristics, students' family academic resources,[32] and schools' racial and free/reduced lunch composition—increased from 1992 to 2013.[33] Thus, the estimated changes in state test scores are based on scores adjusted for student and school demographic differences among states. In addition, the overall test score gain for each state is adjusted for the changing student and school demography of the U.S. NAEP sample. Over this period the U.S. NAEP sample has become progressively poorer, and the share of disadvantaged minority students has grown. Thus, the average adjusted gain in 8th grade math scores across the total sample is greater than the change in observed scores. We chose 8th grade mathematics for our comparison of state gains over this longer period because our overall analysis focuses on mathematics achievement on the PISA, TIMSS, and NAEP tests. However, as noted previously, the gains by states across grades (8th and 4th) and subjects (mathematics and reading) do vary somewhat (see Tables 6A1, 6A2, 6B1, 6B2).[34]

The states in Table 7 are ranked by the average annual point gain in 8th grade math scores between the first time the state participated in NAEP and 2013—in most cases, between 1992 and 2013 (21 years), in some cases between 1996 and 2013 (17 years), and in some cases between 2000 and 2013 (13 years).[35] The takeaways on state performance from this ranking are, not surprisingly, quite similar to the conclusions we drew from Tables 6A1, 6A2, 6B1, and 6B2 for 2003–2013. Students in a group of initially high-scoring, mainly Midwestern/mountain states made very small gains in the 1990s and 2000s. Another group of initially mid- and lower-scoring states—such as Connecticut, California,

and Michigan—also made relatively smaller gains (1 point or less per year). At the other end of the spectrum, some states' students made very large gains: a group of initially high-scoring states (Texas and Vermont), initially low- to middle-scoring states (Delaware, Maryland, Massachusetts, Indiana, and Ohio), initially low-scoring states (Louisiana, North Carolina, and Rhode Island), and one initially very low-scoring state (Hawaii) and the District of Columbia. The argument that it is easier to make large gains when initial scores are lower is in part supported by these results, but both initially lower-scoring and higher-scoring states are represented in the low-gain and high-gain groups. Thus, factors other than the low or high starting point for student math performance should explain why states achieved small or large math gains.

Explaining variation in adjusted state performance

Despite the decline in variance in reported scores after 2009, and the further reduction in variation we obtained by controlling for socioeconomic differences among the states, the remaining state variation in average adjusted scores in Table 7 may be related to state-level variables influencing schools and student performance. These can include spending per student, state-level poverty, teacher union strength, and school accountability, among others. If this were the case, using this information would help us understand further why students in some states score so much higher than students in other states once we take account of variables that explain a large proportion of the differences in observed state average test scores.

It is possible, for example, that the average level of child poverty in a state when these 8th graders were young children—10 years before they took the 8th grade test—is related to the adjusted test score in a state (even though the test score is adjusted for individual student poverty and school-level student poverty). It is possible that students in states with a higher fraction of poor children are less likely to achieve at high levels on the NAEP because of a "low expectations" effect, or a lower cultural capital effect. That is, states with high levels of child poverty may also have lower average levels of ideas and knowledge that raise educational expectations or support student learning in school. Further, strong state accountability systems (Carnoy and Loeb 2003) may be related to higher average state test scores because those states put more emphasis on their students performing well on tests and on constantly improving school performance as measured by those tests. Others have argued that stronger teachers' unions have a negative impact on test scores (Hoxby 1996) and that higher education spending per pupil is not related to student achievement (Hanushek 1986).

We tested whether these variables are related to the residual state 8th grade mathematics test scores shown in Table 7 (as noted previously, these are the state scores adjusted for student and school demographics). We regress the "stacked" test scores on lagged state child poverty, the proportion of the state's population 25 or older with bachelor's degrees, state accountability strength (as calculated in Carnoy and Loeb 2003; Lee and Wong 2004; and Hanushek and Raymond 2005), a measure of union strength from the Current Population Survey,[36] and spending per pupil taken from the *Digest of Educational Statistics* (NCES, various years). We divided the analysis into two periods, 1992–2003 (before the No Child Left Behind Act took effect) and 2003–2013 (after the act took effect). We stack the estimated adjusted state scores over four years of observations (1992, 1996, 2000, 2003) or six years (2003, 2005, 2007, 2009, 2011, 2013) to estimate the influence of the variables just described on state-level test scores over the four or six observed years.[37]

We find that two variables are significantly related to these adjusted state scores: the rate of child poverty in the state 10 years earlier than the 8th grade test date, and the strength of accountability in the state approximately at the time of

the test date. The same two variables are significantly related to average adjusted 8th grade math scores in 1992–2003 and in 2003–2013. Lagged child poverty has a negative relation to state adjusted average test scores, and the coefficient is considerably larger in 1992–2003 than in 2003–2013, -0.39 compared with -0.22. This means that for each 5 percentage-point increase in state-level child poverty (one standard deviation in 1992–2003), test scores decrease in the first period by 2 points (0.25 standard deviations). In the second period, for each increase in state-level child poverty of 6 percentage points (one standard deviation in 2003–2013), average adjusted state test scores decrease by 1.2 points, or about 0.2 standard deviations.

The significant negative relationship between state child poverty and average state test scores (scores that have already been adjusted for individual student race and poverty status, as well as for the degree of concentration of black, Hispanic, and poor students in schools) suggests that there are poverty effects at the state level over and above individual race and social class effects and peer effects at the school level. This implies that students in states with more poverty are likely to have lower achievement whether they are poor or well off, white, black, or Hispanic. This is a very important finding, as it shows that the effect of poverty on education performance is a three-level effect: In addition to the well-documented impact of individual and school-level poverty, state-level poverty puts students at all socioeconomic levels at additional educational disadvantage.

Our other finding is that states that have implemented stronger accountability measures are also likely to have achieved higher adjusted 8th grade math test scores. An increase of 1.4 levels of accountability (one standard deviation) on a scale of 0 to 5 is associated with about one-quarter of a standard deviation increase in average adjusted state test score over 2003–2013. In the earlier period, 1992–2003, the relationship is weaker, with a standard deviation in accountability score (1.28 levels) associated with a 0.15 standard deviation increase in average adjusted test score.

We find no relationship in either period between the adjusted average 8th grade math test score in each state and a state's average expenditures per student in primary and secondary schooling. Rather, it may be that the underlying social conditions or policies associated with higher levels of child poverty are also associated with lower levels of education quality—lower expectations, lower education standards, etc. This holds even when we adjust spending per student for regional price parity, which measures the relative cost of living in each state. A lack of association was also found when examining the relationship between average student performance and the degree of teacher union presence in a state. This finding suggests that the degree of teacher union presence in states provides no additional explanation of student achievement variance across states. Finally, no correlation was found between the proportion of college graduates in a state's adult population and adjusted student performance.

Matchups of states with recently different 8th grade student mathematics score trajectories

We now match and graph the average adjusted and reported 8th grade NAEP mathematics test score trajectories over 1992–2013 for four pairs and one trio of neighboring and/or demographically similar states. It is important to note that these are just examples of state matchups. They were chosen because of geography, demographic similarity, and because they show how neighboring states can differ greatly in their student test score gains over the past 20 years. The four pairs of states are Massachusetts and Connecticut, New York and New Jersey, California and Texas, and Minnesota and Iowa. The trio of neighboring states is Kentucky, Tennessee, and North Carolina.

As shown in **Figure D,** Massachusetts's average 8th grade adjusted mathematics score increased much more than Connecticut's after 2003. In 2003, students in the two states had very similar adjusted and observed scores, but in 2013, adjusted scores were about 0.5 standard deviations higher in Massachusetts than in Connecticut. Similarly, as shown in **Figure E,** New Jersey students' adjusted average score increased more than New York students' after 2003, although the opposite was true before 2003.

Figure F depicts how Texas's average adjusted scores increased slightly more than California's over 1992–2003, and how the gap increased much more rapidly after 2003, despite a recent closing of the gap in 2011–2013. In addition, Texas's adjusted scores were much higher than California's throughout this period.

As shown in **Figure G,** students in the neighboring states of Kentucky, Tennessee, and North Carolina had very similar adjusted 8th grade math scores in 1992. However, North Carolina saw a sharp increase over 1992–2003, so that in 2003, students in North Carolina were scoring 12 to 15 points higher than their counterparts in the other two states. As North Carolina's gains slowed considerably over 2003–2013, Kentucky and Tennessee somewhat closed the gap—first Kentucky, followed by Tennessee over 2011–2013. Thus, after 21 years of change, North Carolina's scores are substantially higher than Kentucky's and Tennessee's.

Figure H shows how in 2000–2003 and again in 2007–2011, Minnesota students made much larger adjusted gains in 8th grade math than did Iowa students. Consequently, by 2013 Minnesota students had opened up a large difference in scores compared with their Iowa counterparts.

Why these different patterns in these matched states? One explanation is that some states undertake successful, system-wide reforms that eventually have larger positive effects on 8th grade students' mathematics scores than in states where reforms are less successful. It is telling that although 8th grade math test scores have risen in all states, they have increased much more in some than in others. Massachusetts apparently was able to enact a combination of changes in its schools, system-wide, in the late 1990s that Connecticut did not. What the state did differently from Connecticut probably influenced its students' mathematics performance.[38] Identifying this difference would be far more productive for policymakers in lower-scoring states than to know what another country did to achieve high test scores sometime in the past. Tanden and Reville (2013) argue that high-scoring Massachusetts managed to make these substantial gains through an improvement effort that began "more than 20 years ago when state legislators passed a major reform law that put rigorous academic goals and well-designed assessments front and center. Importantly, the state also doubled its investment in K-12 education over a seven-year period to help schools and students reach those high standards."

Minnesota apparently made significant changes to its mathematics curriculum in the late 1990s that Iowa did not. This may have been responsible for the greater student performance gains in Minnesota.[39] Similarly, Texas seems to have taken steps in the 1980s and 1990s to raise mathematics scores significantly above California's. Some critics have claimed that Texas excludes a higher fraction of its potential NAEP test takers from the NAEP test than does California. There may be some truth to that, but it does not explain the large difference in math scores between Texas's and California's 8th graders, and particularly the greater increase in test scores in Texas since 2003.

It is worth noting that students in California made somewhat greater gains in the 8th grade and 4th grade reading test over 2003–2013 than students in Texas. Also, students in Connecticut made greater gains in 8th grade reading scores than students in Massachusetts, although students in Massachusetts made larger gains on the 4th grade reading test.

Similarly, students in Iowa made about the same gains on the 8th grade reading test as students in Minnesota, but students in Minnesota made larger gains in 4th grade reading. Kentucky and Tennessee students also made generally somewhat larger gains in both 8th and 4th grade reading test scores over 2003–2013 than students in North Carolina (Tables 6B1 and 6B2), although by 2003 students in North Carolina were not making larger gains in 8th grade mathematics than students in either Kentucky or Tennessee.

Thus, at least in the comparisons between Massachusetts and Connecticut and between Texas and California, whatever the policies were that contributed to the greater gains in mathematics in Texas and Massachusetts did not particularly carry over to improving reading gains, although, adjusted for demographic differences, students in Texas score higher in reading than students in California, and students in Massachusetts score higher in reading than students in Connecticut. There are many reasons why it may have been easier to implement policies that influenced student mathematics performance more than reading performance. Most relevant for this study, however, is the fact that some states were able to have a large impact on how students performed in mathematics—precisely the subject in which U.S. students allegedly perform most poorly on international tests.

Conclusion

For at least two decades, evidence-based policy has been a goal of American education policymakers, who seek data about student knowledge and skills in an effort to use this information to improve schools. One category of such evidence, international test results, has seemingly permitted comparisons of student performance in the United States with that in other countries. Such comparisons have frequently been interpreted to show that American students perform poorly compared with students internationally. From this, reformers conclude that U.S. public education is failing and that its failure imperils America's economic competitiveness.

We have demonstrated that such comparisons are too simplistic and why they need to be used and interpreted with great care. This aligns with the findings of Carnoy and Rothstein (2013). A closer look at the data reveals that disadvantaged FAR students in the United States continue to make very large mathematics and reading gains compared with disadvantaged FAR students in most other countries (Germany and Poland are the exceptions), and that the opposite is true for U.S. advantaged FAR students (Table 2C). Although not quite as extreme, the gains on the TIMSS over 1999–2011 for the United States as a whole follow the same pattern—much greater gains for low FAR groups than for high FAR groups.

We have also shown that international test score interpretations based on U.S. average scores are too simplistic for another reason: The U.S. education system is actually 51 different education systems, as each of the states plus the District of Columbia constitutes its own system. (In fact, there are many more systems when we consider the thousands of school districts, many of them urban districts the size of systems in small countries, that operate quite independently even of their state education administrations.) We show that students in the states vary widely in their performance on the PISA and TIMSS and in their performance gains on the TIMSS. In some states, such as Massachusetts, Minnesota, and North Carolina, students have made large gains in mathematics performance in the 2000s, and these gains have been large for both low and high FAR groups. The gains have also been higher than in Finland, England, and Korea (Tables 5A and 5B).

The wide variation in student test performance in U.S. states (as well as among FAR groups and between subjects described earlier) suggests that U.S. policymakers should reconsider looking to East Asia and some of the European countries for lessons on improving education in lower-performing states. U.S. states that have made large gains and have achieved high performance among all FAR groups are much more likely to provide relevant policy lessons and more pertinent guidance for improving education in states that have made less progress. The contexts for education systems differ among states, but these differences are much smaller than those among lower-performing states on the one hand and, on the other hand, countries with different social, cultural, and educational histories. For example, it makes much more sense for Alabama to look to North Carolina for lessons than to Finland, Poland, or Korea. These reasons, in addition to the others explained in this paper, all illustrate why it is difficult to learn about improving U.S. schools from international test comparisons.

In an attempt to explore what we can learn from student performance in U.S. states over time, we developed a detailed within-country/across-states analysis of NAEP results in mathematics and reading over recent decades. The results of that analysis showed a great variation in how students in various states have progressed over the past decades, with a number of states having made substantial gains in mathematics and reading in the past decade.

In our analysis of how students in various states have progressed over 1992–2013 (Table 7), we showed that students in all states made considerable gains in adjusted 8th grade mathematics test scores, but that the average annual increase in the top-gaining 10 states was 1.6 points, double the annual gain of students in the bottom-gaining 10 states.

For the period 2003–2013, large or small gains were not associated with having either low or high starting levels of student performance. Large gains in mathematics were made, for example, in states starting from a low level of student performance in 2003, such as Hawaii, Rhode Island, Louisiana, and the District of Columbia. Large gains were also made in Texas, North Carolina, Massachusetts, and New Jersey, which started from middle and high levels in 2003. Small gains were made in Alabama and Utah, which started out with low scores, and in Connecticut, Iowa, Michigan, New York, and South Dakota, which started out with middle or high scores.

Also, we learned that states that made large reading gains were not necessarily the same states that made large mathematics gains, and that in some states, mathematics and reading gains were made in one grade, but not in the other. West Virginia 8th graders made small gains in mathematics from a low start, but 4th graders made large gains. California 4th graders made very low gains in mathematics from a low start, but 8th graders made moderate gains. A few high-scoring states such as Massachusetts, Vermont, and New Jersey made large gains in both subjects, and initially low-scoring states such as Hawaii and D.C. made large gains in reading as well as mathematics.

When we explored possible explanations for the variation that remains in average 8th grade mathematics state test scores after adjustments were made, we found that the strength of state school accountability measures is significantly related to adjusted state test scores over both 1992–2003 and 2003–2013. Students in states with stronger accountability systems do better in the NAEP math test, even though that test is not directly linked to the tests that are used to evaluate students within states. Additionally, we tested whether this remaining variation is related to spending per student, and found it is not. We also found that lagged state child poverty rates are significantly related to test scores; students in states with higher child poverty levels perform worse in 8th grade mathematics even when we control for individual

student poverty and average school poverty. In other words, students in states with more poverty are likely to have lower achievement whether they are poor or well off, white, black, or Hispanic.

As a suggestive strategy for further (qualitative) policy research, we paired off neighboring and/or demographically similar states with different patterns of gains in average student performance in 8th grade math. We showed that students in Massachusetts made much larger gains after 2003 than students in neighboring Connecticut; that students in New Jersey made larger gains than students in New York after 2003; that students in Texas already started out scoring higher than students in California in 8th grade math in 1992, but still made larger gains over 1992–2013, especially after 2003; that students in North Carolina made much larger gains over 1992–2003 than students in neighboring Kentucky and Tennessee, but that students first in Kentucky and then in Tennessee caught up somewhat after 2003; and that students in Minnesota made larger gains than students in Iowa almost throughout the entire 20-year period. We argued that each of these comparison groups could provide important insights into the kinds of policies that enabled students in some states to make much larger adjusted gains in math scores than students in neighboring and/or demographically similar states.

In short, instead of looking to foreign countries, U.S. education policymakers should look to U.S. states for the most relevant insights into how to improve student performance. In particular, states with high student performance and substantial performance gains serve as an excellent starting point for this process. We urge researchers and policymakers to make full use of the NAEP results, in combination with suitable qualitative work, to identify the key policies within states that produce such large differences in student achievement trajectories.

Acknowledgments

The authors gratefully acknowledge professors Henry M. Levin, Helen F. Ladd, and David Berliner for their helpful comments and advice on earlier drafts of the paper. They also appreciate Lawrence Mishel's and Elaine Weiss's feedback on versions prior to the completion of the report. The authors are also grateful to Michael McCarthy for having edited this report, to Chris Roof and Tanyell Cooke for their work formatting the tables and figures, and to Alyssa Davis and Will Kimball for their data assistance. Finally, they appreciate the assistance of the Economic Policy Institute communications staff who helped to disseminate it.

About the authors

Martin Carnoy is Vida Jacks Professor of Education and Economics at Stanford University and a research associate of the Economic Policy Institute. He has written more than 30 books and 150 articles on political economy, educational issues, and labor economics. He holds an electrical engineering degree from Caltech and a Ph.D. in economics from the University of Chicago. Much of his work is comparative and international. His recent books include *Sustaining the New Economy: Work, Family and Community in the Information Age*, *The Charter School Dust-Up* (coauthored with Richard Rothstein), *Vouchers and Public School Performance*, *Cuba's Academic Advantage*, and *The Low Achievement Trap*.

Emma García joined the Economic Policy Institute as an economist in 2013. She specializes in the economics of education and education policy. Her areas of research include analysis of the production of education, returns to education, program evaluation, international comparative education, human development, and cost-effectiveness and cost-benefit analysis in education. Prior to joining EPI, García conducted research for the Center for Benefit-Cost Studies of Edu-

cation and other research centers at Teachers College, Columbia University, and did consulting work for MDRC, the Inter-American Development Bank, and the National Institute for Early Education Research. García has a Ph.D. in Economics and Education from Teachers College, Columbia University.

Tatiana Khavenson is a researcher at the Institute of Education at National Research University Higher School of Economics in Moscow. She is a graduate of the sociology department at Moscow State University. Her research interests are studying the factors that influence students' performance in school, and understanding the social sources of educational inequalities, including the social structures that influence students' educational trajectories. Her most recent publications include *Is Brazilian education improving? Evidence from PISA and SAEB; Using TIMSS and PISA results to inform educational policy: A study of Russia and its neighbours; Relationship between the Unified State Exam and higher education performance;* and *Teacher characteristics and student achievements in TIMSS: Findings gained from applying the "first-difference" method to TIMSS-2007 data*.

Tables and figures

TABLE 1

Average national PISA scores, reading and math, United States and eight comparison countries, 2012

		Reading	Math
Top-scoring countries	*Canada*	523	518
	Finland	524	519
	Korea	536	554
	*Average**	528	530
Similar post-industrial countries	*France*	505	495
	Germany	508	514
	U.K.	499	494
	*Average**	504	501
High-scoring newcomer countries	*Poland*	518	518
	Ireland	523	501
	*Average**	521	509
United States, overall and selected states	*U.S. average*	498	481
	Massachusetts	527	514
	Connecticut	521	506
	Florida	492	467
U.S. versus:	*Top-scoring average*	-30	-49
	Similar post-industrial average	-7	-19
	High-scoring newcomer average	-23	-28

* Unweighted average of these countries

Source: EPI analysis of OECD PISA International Database

ECONOMIC POLICY INSTITUTE

TABLE 2A

PISA reading scores, U.S., selected states, and comparison countries, 2012

		Family academic resource (FAR) group							
		Group 1 (Lowest)	Group 2	Group 3	Group 4	Group 5	Group 6	Average with own FAR weights	Average with U.S. FAR weights
					United States				
U.S. overall	*FAR distribution*	22.2%	18.0%	30.0%	14.4%	11.0%	4.3%		
	Average score	448	477	508	536	554	530	498*	499*
Florida	*FAR distribution*	25.2%	23.1%	27.9%	12.7%	7.5%	3.6%		
	Average score	452	482	507	529	539	535	492	498
Connecticut	*FAR distribution*	14.1%	15.0%	28.5%	18.3%	16.4%	7.8%		
	Average score	448	472	522	541	584	588	521	509
Massachusetts	*FAR distribution*	15.9%	16.1%	30.0%	16.4%	15.0%	6.7%		
	Average score	460	485	529	558	588	576	527	519
					Similar post-industrial countries				
France	*FAR distribution*	17.7%	18%	27.5%	17.2%	12.8%	6.8%		
	Average score	428	470	511	546	577	581	505	501
Germany	*FAR distribution*	9.6%	13.3%	28.3%	20.1%	18%	10.7%		
	Average score	430	469	506	530	565	560	508	495
U.K.	*FAR distribution*	14.5%	16.5%	29%	17.8%	14.7%	7.3%		
	Average score	425	465	498	531	561	561	499	490
					Top-scoring countries				
Canada	*FAR distribution*	10.4%	14.8%	31.4%	20.1%	15.3%	8.1%		
	Average score	462	495	520	546	565	567	523	513
Finland	*FAR distribution*	8.7%	13.9%	34.9%	21.4%	15.8%	5.4%		
	Average score	462	486	518	543	574	573	524	512
Korea	*FAR distribution*	4.6%	8.3%	26.9%	23.3%	25.5%	11.4%		
	Average score	464	491	522	537	557	582	536	512
					High-scoring newcomer countries				
Poland	*FAR distribution*	11.2%	20.0%	33.7%	17.3%	10.9%	6.9%		
	Average score	459	490	519	537	568	569	518	511
Ireland	*FAR distribution*	13.6%	16.0%	29.1%	19.5%	14.8%	7.0%		
	Average score	450	487	517	555	576	581	523	512

* Slight discrepancy is an artifact of rounding.

Source: EPI analysis of OECD PISA International Database

ECONOMIC POLICY INSTITUTE

TABLE 2B

PISA mathematics scores, U.S., selected states, and comparison countries, 2012

		Family academic resource (FAR) group							
		Group 1 (Lowest)	Group 2	Group 3	Group 4	Group 5	Group 6	Average with own FAR weights	Average with U.S. FAR weights
					United States				
U.S. overall	*FAR distribution*	22.2%	18.0%	30.0%	14.4%	11.0%	4.3%		
	Average score	432	459	491	513	542	525	481*	482*
Florida	*FAR distribution*	25.2%	23.1%	27.9%	12.7%	7.5%	3.6%		
	Average score	427	450	479	506	532	511	467	473
Connecticut	*FAR distribution*	14.1%	15.0%	28.5%	18.3%	16.4%	7.8%		
	Average score	434	453	505	525	569	580	506	493
Massachusetts	*FAR distribution*	15.9%	16.1%	30.0%	16.4%	15.0%	6.7%		
	Average score	442	468	515	544	576	580	514	504
					Similar post-industrial countries				
France	*FAR distribution*	17.7%	18%	27.5%	17.2%	12.8%	6.8%		
	Average score	423	460	498	533	568	563	495	490
Germany	*FAR distribution*	9.6%	13.3%	28.3%	20.1%	18%	10.7%		
	Average score	430	471	509	536	566	573	514	498
U.K.	*FAR distribution*	14.5%	16.5%	29%	17.8%	14.7%	7.3%		
	Average score	428	462	488	519	556	558	494	485
					Top-scoring countries				
Canada	*FAR distribution*	10.4%	14.8%	31.4%	20.1%	15.3%	8.1%		
	Average score	463	489	516	537	557	558	518	509
Finland	*FAR distribution*	8.7%	13.9%	34.9%	21.4%	15.8%	5.4%		
	Average score	463	485	513	535	561	576	519	508
Korea	*FAR distribution*	4.6%	8.3%	26.9%	23.3%	25.5%	11.4%		
	Average score	474	493	534	554	582	614	554	525
					High-scoring newcomer countries				
Poland	*FAR distribution*	11.2%	20.0%	33.7%	17.3%	10.9%	6.9%		
	Average score	462	483	514	542	574	577	518	510
Ireland	*FAR distribution*	13.6%	16.0%	29.1%	19.5%	14.8%	7.0%		
	Average score	434	469	498	529	547	558	501	491

* Slight discrepancy is an artifact of rounding.

Source: EPI analysis of OECD PISA International Database

ECONOMIC POLICY INSTITUTE

TABLE 2C

PISA reading and mathematics score gap changes, United States versus comparison countries, 2000–2012

Reading

	Gap changes, U.S. versus:							
Family academic resource group	**France**	**Germany**	**U.K.**	**Canada**	**Finland**	**Korea**	**Poland**	**Ireland**
Group 1 (lowest)	+31	-40	+44	+34	+65	+29	-8	+49
Group 2	+16	-44	+27	+18	+50	+21	-23	+21
Group 3	+1	-32	+19	+11	+25	+5	-45	+0
Group 4	-12	-20	+16	+4	+23	+4	-30	-7
Group 5	-26	-31	+2	-8	-2	-14	-52	-16
Group 6 (highest)	-64	-42	-15	-35	-23	-57	-81	-45

Mathematics

	Gap changes, U.S. versus:							
Family academic resource group	**France**	**Germany**	**U.K.**	**Canada**	**Finland**	**Korea**	**Poland**	**Ireland**
Group 1 (lowest)	+51	-34	+46	+40	+60	+15	-29	+41
Group 2	+37	-41	+34	+26	+46	+21	-28	+13
Group 3	+20	-37	+32	+8	+16	+6	-48	-3
Group 4	+7	-34	+24	+5	+11	+5	-56	-14
Group 5	-11	-30	+6	-7	-3	-7	-68	-9
Group 6 (highest)	-48	-53	-8	-32	-40	-55	-93	-49

Note: Numbers in this table take the 2012 U.S. average score for a FAR group, less the 2012 comparison country's average score for the same FAR group, and subtract from this result the 2000 U.S. average score for that FAR group, less the 2000 comparison country's average score for the same FAR group.

Source: EPI analysis of OECD PISA International Database

ECONOMIC POLICY INSTITUTE

TABLE 3A

TIMSS mathematics score averages, United States and comparison countries/provinces, 2011

		Family academic resource (FAR) group						
		Group 1 (lowest)	Group 2	Group 3	Group 4	Group 5/6 (higher/ highest)	Average	Average with U.S. FAR weights
U.S.	*FAR distribution*	15.9%	22.6%	28.2%	17.5%	15.8%		
	Average score	465	485	516	542	548	509*	510*
Canadian provinces								
Ontario	*FAR distribution*	7.8%	17.9%	33.3%	20.9%	20.0%		
	Average score	471	482	509	526	538	512	504
Alberta	*FAR distribution*	7.9%	17.2%	31.8%	21.6%	21.6%		
	Average score	474	484	505	516	525	505	500
Quebec	*FAR distribution*	14.8%	25.3%	33.2%	15.1%	11.6%		
	Average score	502	514	539	552	563	532	534
Comparison countries								
Finland	*FAR distribution*	7.2%	17.1%	34.7%	21.8%	19.3%		
	Average score	465	493	514	530	535	514	507
Korea	*FAR distribution*	8.6%	10.3%	25.3%	23.7%	32.1%		
	Average score	546	556	594	627	653	613	593
England	*FAR distribution*	16.4%	22.3%	28.1%	17.0%	16.2%		
	Average score	442	481	518	540	556	507	507

* Slight discrepancy is an artifact of rounding.

Source: EPI analysis of TIMSS International Database

ECONOMIC POLICY INSTITUTE

TABLE 3B

TIMSS mathematics score averages, U.S. overall and participating states, 2011

		Family academic resource (FAR) group						
		Group 1 (lowest)	Group 2	Group 3	Group 4	Group 5/6 (higher/ highest)	Average	Average with U.S. FAR weights
U.S.	*FAR distribution*	15.9%	22.6%	28.2%	17.5%	15.8%		
	Average score	465	485	516	542	548	509*	510*
Massachusetts	*FAR distribution*	10.8%	15.7%	27.5%	21.9%	24.0%		
	Average score	503	522	563	575	598	561	552
Minnesota	*FAR distribution*	10.1%	16.1%	30.1%	20.5%	23.3%		
	Average score	494	506	543	568	574	545	536
Connecticut	*FAR distribution*	11.7%	18.3%	27.0%	20.2%	22.8%		
	Average score	446	475	521	550	565	518	511
Indiana	*FAR distribution*	14.1%	21.5%	32.1%	16.6%	15.7%		
	Average score	479	500	526	544	558	522	521
Alabama	*FAR distribution*	25.1%	26.1%	25.5%	12.8%	10.4%		
	Average score	434	448	481	510	502	466	474
Colorado	*FAR distribution*	15.0%	17.9%	28.2%	17.4%	21.5%		
	Average score	464	487	521	544	557	518	514
North Carolina	*FAR distribution*	15.6%	21.9%	29.1%	16.9%	16.5%		
	Average score	484	518	539	560	585	537	537
California	*FAR distribution*	18.0%	28.0%	29.0%	14.0%	11.0%		
	Average score	452	469	507	532	535	493	499
Florida	*FAR distribution*	19.0%	24.6%	29.6%	15.0%	11.8%		
	Average score	484	498	518	544	553	513	518

* Slight discrepancy is an artifact of rounding.

Source: EPI analysis of TIMSS International Database

ECONOMIC POLICY INSTITUTE

TABLE 4

8th grade TIMSS math trends, select U.S. states, 1995–2011

State	1995	1999	2007	2011	Average annual change
Connecticut	—	512	—	518	+0.1%
Massachusetts	—	513	547	561	+0.7%
Minnesota	518	—	532	545	+0.3%
North Carolina	—	495	—	537	+0.7%
Indiana	—	515	—	522	+0.1%
Missouri	505	490	—	—	-0.8%
Oregon	525	514	—	—	-0.5%

Source: Harmon et al. (1997); Mullis et al. (1998); Mullis et al. (2001a); Mullis et al. (2012); NCES NAEP Data Explorer

ECONOMIC POLICY INSTITUTE

TABLE 5A

Changes in 8th grade TIMSS mathematics scores, U.S. and selected states and countries, 1999–2011

		Family academic resource (FAR) group						
		Group 1 (lowest)	Group 2	Group 3	Group 4	Group 5/6 (higher/ highest)	Average	Average with 2011 U.S. FAR weights
					United States			
U.S.	*1999*	439	461	495	523	537	502	490
	2011	465	485	516	542	548	509*	510*
	Difference	26	24	21	20	10	8	21
Connecticut	*1999*	436	461	507	519	544	512	493
	2011	446	475	521	550	565	518	511
	Difference	10	14	14	31	21	6	17
Massachusetts	*1999*	445	476	508	518	543	513	498
	2011	503	522	563	575	598	561	552
	Difference	58	45	55	57	55	48	54
Indiana	*1999*	438	483	504	521	531	515	496
	2011	479	500	526	544	558	522	521
	Difference	42	17	21	23	27	7	25
North Carolina	*1999*	439	463	486	517	525	495	485
	2011	484	518	539	560	585	537	537
	Difference	46	55	54	43	60	42	52
					Comparison countries			
Finland	*1999*	483	492	521	527	538	520	512
	2011	465	493	514	530	535	514	507
	Difference	-18	1	-7	2	-2	-6	-5
Korea	*1999*	527	550	581	605	625	587	577
	2011	546	556	594	627	653	613	593
	Difference	19	6	13	22	29	26	16
England	*1999*	438	456	488	505	537	496	484
	2011	442	481	518	540	556	507	507
	Difference	5	25	30	35	19	10	24

* Slight discrepancy is an artifact of rounding.

Source: EPI analysis of TIMSS International Database

ECONOMIC POLICY INSTITUTE

TABLE 5B

Changes in 8th grade TIMSS mathematics scores, U.S., Minnesota, Korea, and England, 1995–2011

		Family academic resource (FAR) group						
		Group 1 (lowest)	Group 2	Group 3	Group 4	Group 5/6 (higher/ highest)	Average	Average with 2011 U.S. FAR weights
U.S.	*1995*	412	435	474	499	513	492	466
	2011	465	485	516	542	548	509*	510*
	Difference	53	50	42	43	35	17	45
Minnesota	*1995*	444	464	499	515	534	518	491
	2011	494	506	543	568	574	545	536
	Difference	49	42	44	53	39	27	45
Korea	*1995*	505	523	562	592	609	581	557
	2011	546	556	594	627	653	613	593
	Difference	41	33	32	35	45	32	36
England	*1995*	398	435	474	498	521	498	465
	2011	442	481	518	540	556	507	507
	Difference	44	46	45	42	35	9	43

* Slight discrepancy is an artifact of rounding.

Source: EPI analysis of TIMSS International Database

ECONOMIC POLICY INSTITUTE

TABLE 6A1

8th grade mathematics adjusted NAEP scores, by state, 2003, 2007, 2011, 2013

2003		2007		2011		2013	
Hawaii	260.5	Hawaii	261.9	West Virginia	273.9	Alabama	277.7
Nevada	269.1	Alabama	271.4	Alabama	274.3	West Virginia	279.5
Alabama	271.0	D.C.	271.8	Hawaii	274.4	Michigan	280.2
West Virginia	272.0	Rhode Island	272.3	California	275.4	Utah	280.8
Rhode Island	272.4	Nevada	272.8	Utah	276.2	Connecticut	281.0
California	272.6	West Virginia	273.2	Michigan	276.7	California	281.2
D.C.	272.9	Michigan	274.7	Tennessee	276.8	Hawaii	281.4
Arkansas	273.0	Utah	275.3	Nebraska	278.5	New York	283.2
Mississippi	273.1	Connecticut	275.6	Connecticut	279.1	Iowa	283.7
Tennessee	273.8	New Mexico	277.0	New York	279.3	Wyoming	283.9
Michigan	274.5	Tennessee	277.0	Nevada	279.6	Rhode Island	284.3
Utah	274.7	California	277.0	Rhode Island	280.4	Oklahoma	284.9
Kentucky	276.1	Oklahoma	277.3	Wyoming	280.6	Idaho	285.2
Oklahoma	276.1	Arizona	278.6	Missouri	281.0	Nebraska	285.3
New Mexico	276.3	New Hampshire	278.8	D.C.	281.4	New Mexico	285.3
Louisiana	276.4	Mississippi	279.0	Mississippi	281.5	Tennessee	285.3
Pennsylvania	276.7	Arkansas	279.2	Iowa	281.7	Kentucky	285.5
New Hampshire	276.8	Nebraska	279.5	Arizona	282.1	Nevada	285.5
Maine	277.2	Illinois	280.0	Oklahoma	282.4	South Dakota	286.3
Idaho	277.5	Wisconsin	280.4	New Mexico	282.5	North Dakota	286.4
Nebraska	277.5	Georgia	280.5	Oregon	282.7	Arizona	286.5
Maryland	277.5	Kentucky	280.6	Kentucky	283.1	Mississippi	286.8
Wyoming	277.9	Wyoming	280.9	New Hampshire	283.4	Oregon	287.4
North Dakota	277.9	New York	281.0	Florida	283.5	Missouri	288.1
Colorado	277.9	Iowa	281.7	Idaho	284.0	Arkansas	288.3
Arizona	278.2	Colorado	281.8	Indiana	284.4	Illinois	288.5
Massachusetts	278.7	Missouri	281.9	Louisiana	284.5	Montana	288.5

TABLE 6A1 (CONTINUED)

2003		2007		2011		2013	
New Jersey	278.8	Ohio	282.3	Illinois	284.6	New Hampshire	289.1
Illinois	278.8	Maine	282.4	South Dakota	284.7	Virginia	289.2
Ohio	278.9	Florida	282.9	Georgia	284.8	Colorado	289.2
Florida	278.9	North Dakota	283.0	North Dakota	285.1	Wisconsin	289.3
New York	279.6	New Jersey	283.0	Maryland	285.3	Delaware	289.7
Missouri	279.6	South Dakota	283.2	Wisconsin	285.4	Georgia	289.7
Washington	279.7	Oregon	283.3	Pennsylvania	285.7	South Carolina	290.2
Oregon	279.7	Idaho	283.3	Arkansas	285.8	Florida	290.3
Wisconsin	279.7	Washington	283.4	Washington	285.8	D.C.	290.4
Connecticut	279.8	Indiana	283.6	Virginia	285.9	Pennsylvania	290.8
Indiana	279.8	Maryland	284.0	Maine	286.3	Louisiana	290.8
Iowa	279.9	Louisiana	284.0	Colorado	286.3	Minnesota	291.0
Vermont	280.0	Pennsylvania	284.0	South Carolina	286.6	Washington	291.2
Delaware	280.0	Minnesota	284.3	Ohio	287.2	Maine	291.2
Virginia	280.1	Montana	284.7	Kansas	287.5	Maryland	291.3
South Dakota	280.3	Vermont	285.0	New Jersey	287.8	Kansas	292.0
Georgia	280.3	Virginia	285.2	Delaware	288.3	Indiana	292.0
Montana	280.4	Kansas	286.7	Minnesota	289.0	New Jersey	292.4
Kansas	281.3	Delaware	286.8	Montana	289.5	Ohio	293.1
Texas	282.8	South Carolina	288.3	Vermont	290.1	Vermont	295.6
Minnesota	284.5	North Carolina	288.8	North Carolina	291.6	North Carolina	296.1
South Carolina	284.9	Massachusetts	289.6	Massachusetts	292.6	Texas	297.0
North Carolina	286.8	Texas	290.6	Texas	296.5	Massachusetts	298.0

Note: Scores are adjusted for student characteristics, student family background, school socioeconomic status and race composition, and teacher characteristics.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

TABLE 6A2

4th grade mathematics adjusted NAEP scores, by state, 2003, 2007, 2011, 2013

2003		2007		2011		2013	
Hawaii	218.7	Hawaii	225.4	Alabama	233.1	California	234.9
Delaware	224.8	Delaware	230.9	Michigan	233.5	Alabama	235.2
Nevada	225.8	Nevada	231.4	West Virginia	234.3	Connecticut	235.9
Utah	228.5	Rhode Island	232.3	Delaware	235.4	Michigan	236.4
Alabama	229.0	Arizona	232.7	Tennessee	235.4	Utah	236.5
Kentucky	229.1	California	233.9	Hawaii	235.5	South Dakota	236.9
Tennessee	229.6	Tennessee	233.9	Arizona	235.9	Hawaii	237.2
Colorado	229.7	Alabama	234.1	New York	236.2	Idaho	237.4
Arizona	230.1	Utah	234.1	Oregon	236.2	New York	237.8
Rhode Island	230.2	Michigan	234.2	California	236.3	Illinois	238.5
West Virginia	230.7	Oregon	235.0	Connecticut	236.4	North Dakota	238.5
California	231.0	Connecticut	235.4	Idaho	236.7	New Mexico	238.6
Idaho	231.4	New Mexico	235.6	Utah	236.8	Rhode Island	238.6
North Dakota	231.7	Nebraska	236.6	Nebraska	237.0	Nevada	239.1
Maryland	231.7	Kentucky	236.7	Rhode Island	237.2	Iowa	239.5
Montana	231.8	West Virginia	237.0	South Dakota	238.0	Missouri	239.7
New Jersey	232.0	Colorado	237.5	Mississippi	238.2	Arizona	240.0
Wisconsin	232.0	Illinois	237.5	Illinois	238.2	Nebraska	240.1
Oklahoma	232.1	Idaho	237.6	Iowa	238.6	Montana	240.2
Maine	232.3	Oklahoma	237.9	New Mexico	238.8	Oregon	240.7
Nebraska	232.4	South Dakota	238.1	Wyoming	238.8	Pennsylvania	240.7
South Dakota	232.5	Mississippi	238.3	North Dakota	239.3	Wisconsin	240.9
Illinois	232.5	Maryland	238.4	Oklahoma	239.4	Mississippi	241.0
New Mexico	232.6	Iowa	238.4	Nevada	239.6	West Virginia	241.0
Arkansas	232.9	New Hampshire	238.6	Louisiana	240.0	Louisiana	241.2
Michigan	233.2	Maine	238.9	Montana	240.0	Tennessee	241.3
Iowa	233.2	Washington	239.1	Kentucky	240.7	Wyoming	241.3
New Hampshire	233.2	Missouri	239.2	Maine	240.7	Kentucky	241.4
Mississippi	233.3	Wyoming	239.3	Missouri	240.8	Oklahoma	241.5

TABLE 6A2 (CONTINUED)

2003		2007		2011		2013	
Oregon	233.6	Georgia	239.5	Washington	241.0	Virginia	241.6
Connecticut	233.7	Wisconsin	239.7	Georgia	241.0	New Jersey	241.7
Pennsylvania	233.8	North Dakota	239.9	Virginia	241.1	Colorado	241.8
Georgia	234.0	Minnesota	240.2	Wisconsin	241.3	Washington	242.1
Missouri	234.0	Montana	240.7	Colorado	241.4	South Carolina	242.2
Indiana	234.2	New Jersey	240.8	New Jersey	241.4	New Hampshire	242.6
Minnesota	234.3	Vermont	240.8	Pennsylvania	241.6	Arkansas	243.1
Vermont	234.3	Virginia	240.9	Vermont	241.7	Ohio	243.2
Washington	234.6	Pennsylvania	241.0	New Hampshire	241.7	Vermont	243.3
Ohio	235.3	Louisiana	241.0	D.C.	241.7	Maine	243.3
New York	235.8	New York	241.1	South Carolina	241.8	Delaware	244.1
Wyoming	235.9	Arkansas	241.2	Arkansas	242.1	Georgia	244.1
Massachusetts	235.9	Ohio	241.7	Ohio	242.3	D.C.	244.8
D.C.	236.1	South Carolina	242.0	Indiana	243.6	Kansas	245.0
Virginia	236.3	D.C.	242.8	Florida	244.6	Maryland	245.2
Kansas	238.3	Florida	243.9	Maryland	244.9	Florida	245.4
Louisiana	238.4	Indiana	244.0	Kansas	244.9	Massachusetts	246.6
Florida	239.0	Massachusetts	244.6	Minnesota	245.1	Minnesota	247.5
Texas	241.0	North Carolina	244.6	Texas	245.7	Texas	247.9
South Carolina	241.6	Kansas	244.8	Massachusetts	247.0	North Carolina	248.1
North Carolina	244.4	Texas	246.0	North Carolina	247.5	Indiana	249.2

Note: Scores are adjusted for student characteristics, student family background, school socioeconomic status and race composition, and teacher characteristics.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

TABLE 6B1

8th grade reading adjusted NAEP scores, by state, 2003, 2007, 2011, 2013

2003		2007		2011		2013	
Hawaii	251.7	Hawaii	248.7	Hawaii	255.7	West Virginia	260.4
Nevada	254.2	Nevada	254.1	West Virginia	257.4	Hawaii	262.4
Utah	257.9	West Virginia	256.4	California	258.6	North Dakota	263.5
California	259.9	Arizona	256.5	D.C.	260.6	Alabama	263.9
Alabama	260.3	Rhode Island	256.8	Nevada	260.7	Mississippi	264.7
West Virginia	260.5	Alabama	256.9	Utah	260.8	Arizona	265.6
Tennessee	260.6	D.C.	257.2	North Dakota	261.0	Utah	266.3
Maryland	261.2	Utah	257.3	Arizona	261.5	South Dakota	266.8
D.C.	261.4	Tennessee	258.1	Tennessee	261.9	Michigan	266.9
Arkansas	262.0	California	258.1	Iowa	262.3	California	267.0
Minnesota	262.1	New Mexico	258.3	Michigan	262.4	Rhode Island	267.3
North Dakota	262.2	Michigan	258.6	Alabama	262.9	Iowa	267.5
Iowa	262.2	Wisconsin	259.7	South Carolina	263.3	New Mexico	267.7
New Mexico	262.3	North Dakota	260.0	New Mexico	263.3	Wyoming	267.9
New Hampshire	262.4	New Hampshire	260.8	Mississippi	263.3	New Hampshire	268.0
Arizona	262.5	Mississippi	260.9	New Hampshire	263.5	Nevada	268.2
Michigan	262.5	Kentucky	261.0	Rhode Island	263.9	Kansas	268.5
Washington	262.5	Wyoming	261.1	South Dakota	264.0	Wisconsin	268.5
Idaho	262.6	Washington	261.1	Wisconsin	264.5	Minnesota	268.5
Wisconsin	262.7	Connecticut	261.3	Oklahoma	264.6	Virginia	268.5
Louisiana	262.7	Minnesota	261.5	Nebraska	264.9	Idaho	268.8
Oregon	262.9	Colorado	262.1	Oregon	265.0	New York	268.8
Colorado	263.1	South Carolina	262.4	Arkansas	265.2	D.C.	268.8
Wyoming	263.2	Arkansas	262.4	New York	265.2	Oklahoma	268.9
Pennsylvania	263.5	Indiana	262.6	Texas	265.3	Nebraska	269.0
Maine	263.5	Oklahoma	262.7	Virginia	265.3	Tennessee	269.4
Georgia	263.7	Iowa	262.7	Indiana	265.4	Indiana	269.5

TABLE 6B1 (CONTINUED)

2003		2007		2011		2013	
Ohio	263.8	Idaho	262.8	Louisiana	265.5	South Carolina	269.5
Connecticut	263.8	Illinois	262.8	Wyoming	265.5	Maine	269.7
Kansas	263.8	Nebraska	262.9	Colorado	265.6	Arkansas	269.8
Indiana	263.9	Georgia	263.1	Idaho	265.8	Delaware	270.0
Rhode Island	264.0	South Dakota	263.2	Washington	265.9	Montana	270.3
Delaware	264.2	North Carolina	263.3	Georgia	265.9	Missouri	270.5
South Carolina	264.3	Louisiana	263.3	Kansas	266.1	Connecticut	270.6
Mississippi	264.3	Florida	263.4	Minnesota	266.3	Illinois	270.8
Nebraska	264.3	Maryland	263.4	Maryland	266.3	Colorado	270.8
Florida	264.9	Missouri	263.6	Delaware	266.6	Texas	270.8
North Carolina	265.0	Virginia	263.8	Pennsylvania	266.8	Louisiana	271.0
South Dakota	265.0	Kansas	264.1	North Carolina	266.8	Ohio	271.4
Oklahoma	265.3	New York	264.1	Florida	267.1	Georgia	271.4
Missouri	265.3	Oregon	264.3	Missouri	267.6	Oregon	271.4
Texas	265.6	Maine	264.6	Illinois	267.7	North Carolina	271.7
Montana	265.8	Texas	265.2	Ohio	267.9	Florida	272.3
Kentucky	265.8	Pennsylvania	265.6	Montana	268.6	Pennsylvania	272.3
Virginia	265.8	New Jersey	266.2	Kentucky	268.7	Washington	272.3
New Jersey	266.3	Ohio	266.4	Connecticut	268.9	Maryland	272.5
Vermont	266.4	Delaware	266.5	New Jersey	269.0	Kentucky	272.7
Illinois	267.6	Vermont	266.5	Maine	269.7	New Jersey	273.1
New York	268.5	Montana	266.7	Vermont	270.7	Vermont	273.8
Massachusetts	268.5	Massachusetts	267.5	Massachusetts	271.5	Massachusetts	274.3

Note: Scores are adjusted for student characteristics, student family background, school socioeconomic status and race composition, and teacher characteristics.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

TABLE 6B2

4th grade reading adjusted NAEP scores, by state, 2003, 2007, 2011, 2013

2003		2007		2011		2013	
Hawaii	202.2	Hawaii	204.0	Hawaii	209.9	Hawaii	208.7
Nevada	204.5	Nevada	210.0	D.C.	211.9	Arizona	211.9
D.C.	208.6	Arizona	210.3	Utah	212.9	New Mexico	213.3
Arizona	210.5	Oregon	212.8	Arizona	214.3	South Dakota	213.8
California	210.6	D.C.	213.0	Michigan	214.3	Idaho	214.6
Utah	212.2	Utah	213.9	New Mexico	214.3	Utah	214.6
Alabama	213.0	Tennessee	214.4	West Virginia	214.5	North Dakota	214.7
Oregon	213.5	California	214.4	Tennessee	215.2	California	215.4
Wisconsin	213.5	Michigan	214.9	California	215.3	Michigan	216.2
Idaho	213.8	Rhode Island	215.3	South Dakota	215.3	Wisconsin	216.9
Tennessee	213.8	Wisconsin	216.0	Iowa	216.2	Iowa	217.5
Minnesota	213.9	New Hampshire	216.6	Mississippi	216.3	Nevada	217.5
Michigan	214.1	West Virginia	216.6	Wisconsin	216.6	Montana	217.9
North Dakota	214.1	Oklahoma	216.9	North Dakota	216.8	Mississippi	218.6
Rhode Island	215.1	North Dakota	217.4	Idaho	217.1	Illinois	218.8
New Hampshire	215.2	Connecticut	217.5	Nevada	217.1	West Virginia	218.9
Pennsylvania	215.2	Minnesota	217.5	Minnesota	217.3	Nebraska	219.3
Illinois	215.3	Iowa	218.0	Oregon	217.3	Washington	219.5
Indiana	215.3	New Mexico	218.2	South Carolina	217.7	Rhode Island	219.7
Iowa	215.7	Mississippi	218.2	Rhode Island	218.3	Wyoming	219.7
Mississippi	215.7	South Dakota	218.3	Wyoming	218.3	South Carolina	219.7
Wyoming	216.0	Illinois	218.5	Washington	218.5	Oregon	219.8
Maryland	216.0	South Carolina	218.5	Montana	218.6	Alabama	220.1
New Jersey	216.1	Wyoming	218.7	Maine	218.6	New Hampshire	220.4
Kansas	216.2	Indiana	218.7	Oklahoma	218.7	Colorado	220.6
New Mexico	216.4	Nebraska	218.7	Missouri	218.8	Oklahoma	220.6
Colorado	216.5	Idaho	219.0	Illinois	218.9	Tennessee	220.6
Ohio	216.6	Washington	219.2	New Hampshire	219.0	Ohio	220.8
Nebraska	216.7	Louisiana	219.5	Nebraska	219.1	Connecticut	221.0

TABLE 6B2 (CONTINUED)

2003		2007		2011		2013	
Louisiana	216.9	New Jersey	219.5	Virginia	219.9	Minnesota	221.2
Arkansas	216.9	Kansas	219.6	Connecticut	220.1	Missouri	221.7
Maine	217.0	Maine	219.9	Louisiana	220.4	New York	222.1
Washington	217.1	Ohio	220.0	Indiana	220.5	Texas	222.4
Kentucky	217.2	Maryland	220.1	Vermont	220.5	Maine	222.5
Oklahoma	217.3	Arkansas	220.1	Colorado	220.5	Kansas	222.7
Vermont	217.3	Colorado	220.3	Alabama	220.7	Arkansas	222.7
Georgia	217.3	Georgia	220.4	Ohio	220.8	Louisiana	222.9
West Virginia	217.4	Vermont	220.4	Georgia	221.0	New Jersey	223.0
Montana	217.5	Missouri	220.5	Arkansas	221.2	Virginia	223.0
South Dakota	217.6	North Carolina	220.6	New York	221.5	Pennsylvania	223.2
Texas	218.3	New York	220.9	New Jersey	221.8	D.C.	223.2
Virginia	218.5	Alabama	220.9	Pennsylvania	221.9	Vermont	223.6
Missouri	219.3	Pennsylvania	221.1	Texas	222.3	Kentucky	223.7
Connecticut	219.9	Montana	222.1	Kansas	222.4	Georgia	224.0
South Carolina	219.9	Kentucky	222.2	North Carolina	222.6	Delaware	225.3
Massachusetts	221.6	Virginia	222.3	Kentucky	223.1	North Carolina	225.5
Delaware	221.8	Texas	222.4	Delaware	224.1	Indiana	225.6
New York	221.9	Delaware	224.2	Maryland	224.8	Massachusetts	226.9
Florida	222.2	Massachusetts	225.3	Massachusetts	228.6	Maryland	227.0
North Carolina	223.3	Florida	225.4	Florida	229.6	Florida	231.1

Note: Scores are adjusted for student characteristics, student family background, school socioeconomic status and race composition, and teacher characteristics.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

TABLE 7

Gains in 8th grade math adjusted NAEP scores, by state, 1992–2013

State	Adjusted scores									2013 minus earliest test year	Annual gain (points/ year)
	1992	1996	2000	2003	2005	2007	2009	2011	2013		
Iowa	273.0	275.9	—	280.7	280.2	283.7	281.6	284.7	283.6	10.6	0.51
South Dakota	—	—	—	280.9	283.5	285.1	287.9	288.1	286.3	5.4	0.54
Connecticut	265.7	271.8	274.3	280.5	278.4	277.7	282.7	282.8	281.5	15.7	0.75
North Dakota	270.8	276.2	273.8	279.2	281.4	286.0	288.8	288.9	286.9	16.1	0.77
Montana	—	274.9	279.5	280.5	283.8	285.8	288.3	292.3	288.1	13.2	0.78
Nebraska	268.6	274.7	274.1	278.1	280.9	282.4	282.8	281.9	285.6	17.0	0.81
Wyoming	266.1	266.9	269.8	279.4	279.9	284.3	281.7	284.4	284.8	18.7	0.89
Arizona	—	269.8	270.1	277.5	279.7	278.7	282.5	283.8	285.1	15.3	0.90
Michigan	262.0	270.8	271.7	275.5	276.4	276.8	278.9	280.7	281.0	19.1	0.91
Wisconsin	268.9	275.9	279.5	280.0	281.1	281.6	286.0	288.3	288.7	19.9	0.95
Illinois	—	—	275.6	279.3	281.7	281.5	284.2	287.7	287.9	12.3	0.95
California	260.3	266.1	265.0	273.5	276.2	278.4	278.8	278.3	280.7	20.4	0.97
Alabama	257.7	263.9	270.0	272.2	272.8	274.3	278.1	278.5	278.3	20.7	0.99
Maine	269.4	274.6	275.2	277.2	279.8	283.2	285.4	288.7	290.5	21.1	1.00
Oklahoma	262.6	—	270.7	276.4	277.0	278.4	280.5	285.1	283.7	21.1	1.01
New Mexico	263.6	272.3	266.8	276.8	275.0	278.9	280.8	285.0	284.7	21.2	1.01
Idaho	263.6	—	272.8	277.5	279.3	284.6	284.2	286.8	284.8	21.2	1.01
Missouri	266.2	270.7	272.2	279.6	278.7	283.4	285.3	283.7	287.5	21.4	1.02
Utah	259.2	264.9	265.7	275.0	274.6	277.1	277.6	279.4	280.5	21.4	1.02
Kansas	—	—	277.4	281.4	283.6	288.1	289.7	290.3	291.2	13.8	1.06
New York	262.0	270.2	274.0	281.1	283.1	283.8	284.9	283.1	284.3	22.3	1.06
Oregon	—	269.4	274.2	279.8	282.6	284.3	286.5	286.1	287.5	18.1	1.06
Nevada	—	266.9	265.0	269.6	273.2	274.2	276.0	282.3	285.3	18.4	1.08
Minnesota	268.8	275.9	279.5	285.3	286.9	286.8	290.8	293.0	291.6	22.8	1.09
Arkansas	263.8	263.4	264.5	274.1	279.0	281.1	284.3	288.2	287.6	23.8	1.13
South Carolina	266.2	270.0	275.8	285.2	290.4	290.3	292.4	289.4	290.0	23.8	1.14
New Hampshire	265.6	271.8	—	277.3	278.2	280.8	286.1	286.7	289.6	24.0	1.14
West Virginia	254.9	261.0	269.4	273.5	273.7	275.9	276.1	277.3	279.3	24.4	1.16
Virginia	264.1	266.9	272.7	280.8	284.8	286.8	286.4	289.0	288.9	24.8	1.18
Colorado	264.3	271.4	—	278.6	280.2	283.7	284.3	289.8	289.2	24.9	1.19
New Jersey	267.4	273.8	—	279.1	283.2	284.7	287.8	290.5	292.4	25.0	1.19
Georgia	264.7	265.9	272.0	280.7	281.1	282.3	286.0	287.8	289.9	25.2	1.20
Kentucky	259.7	266.2	271.1	276.1	277.6	282.7	283.4	286.0	285.4	25.7	1.22
Mississippi	259.6	264.8	264.5	273.5	277.9	280.4	280.0	283.7	285.5	26.0	1.24
Tennessee	257.9	263.5	266.4	273.2	278.0	278.4	280.2	278.9	283.9	26.1	1.24
Washington	—	269.6	—	279.3	282.2	283.6	286.4	288.4	290.8	21.2	1.25
Texas	269.0	274.9	280.6	283.6	288.3	291.7	292.8	298.7	295.6	26.7	1.27
Delaware	261.1	266.1	—	279.9	283.4	287.1	288.9	290.8	288.7	27.6	1.32
Florida	262.1	269.0	—	279.2	282.7	284.1	288.5	286.3	289.9	27.8	1.32
Pennsylvania	263.2	—	—	277.9	280.5	286.2	287.9	289.1	291.2	28.0	1.33
Indiana	263.4	270.4	276.6	281.1	284.0	286.8	287.4	288.4	292.4	29.0	1.38

TABLE 7 (CONTINUED)

State	Adjusted scores 1992	1996	2000	2003	2005	2007	2009	2011	2013	2013 minus earliest test year	Annual gain (points/ year)
Maryland	261.7	268.7	274.2	278.1	278.5	285.2	287.4	287.9	291.2	29.5	1.40
Ohio	263.2	—	275.3	279.1	282.0	284.0	285.8	290.8	293.1	30.0	1.43
Rhode Island	255.2	263.2	268.9	273.9	272.8	275.0	279.5	284.1	285.3	30.1	1.43
Vermont	—	271.1	274.3	280.6	283.1	286.6	289.8	293.3	295.9	24.8	1.46
Louisiana	258.0	261.8	272.9	275.8	281.6	284.8	288.6	286.3	290.1	32.1	1.53
District of Columbia	259.3	260.7	265.2	273.5	275.9	274.5	284.2	285.7	292.2	32.8	1.56
North Carolina	260.2	270.0	279.4	287.6	288.0	290.4	291.6	294.4	296.0	35.9	1.71
Massachusetts	262.1	270.5	274.7	279.6	286.2	291.2	295.2	295.6	298.4	36.3	1.73
Hawaii	241.7	251.9	254.8	260.2	259.6	262.9	267.5	277.1	281.2	39.5	1.88

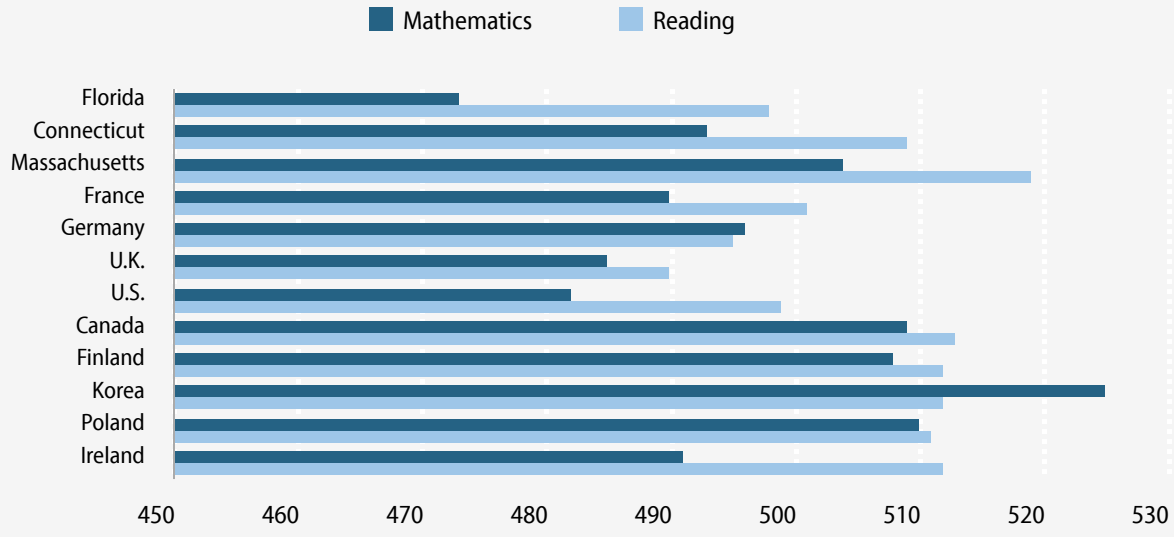
Note: States are ranked by annual gain. Scores are adjusted within each test year for student characteristics, student family socioeconomic (SES) characteristics, and racial and SES composition of school. Test scores are adjusted across years for changing social class composition of sample by using 1996 student characteristics, family academic resources, and school racial and SES composition to weight all years.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

FIGURE A

PISA math and reading scores adjusted to U.S. FAR weights, United States, select states, and comparison countries, 2012

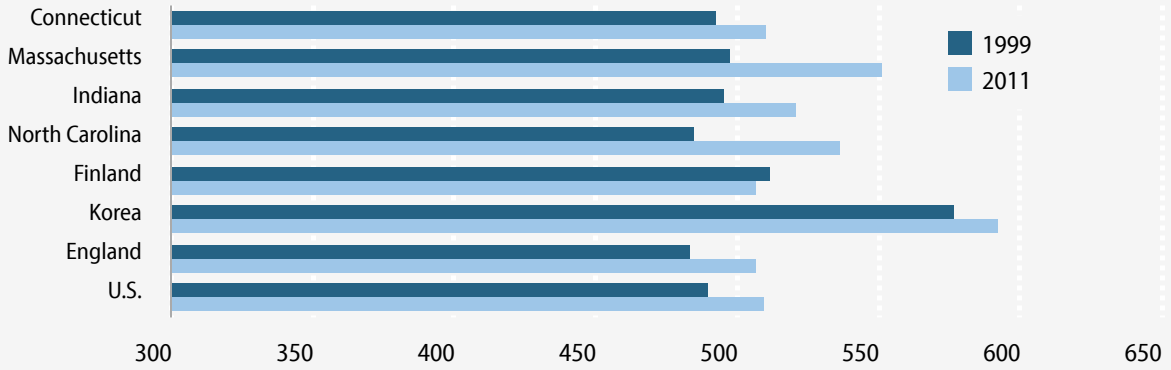


Source: EPI analysis of OECD PISA International Database

ECONOMIC POLICY INSTITUTE

FIGURE B

TIMSS 8th grade math scores adjusted to 2011 U.S. FAR weights, United States, select states, and select countries, 1999–2011

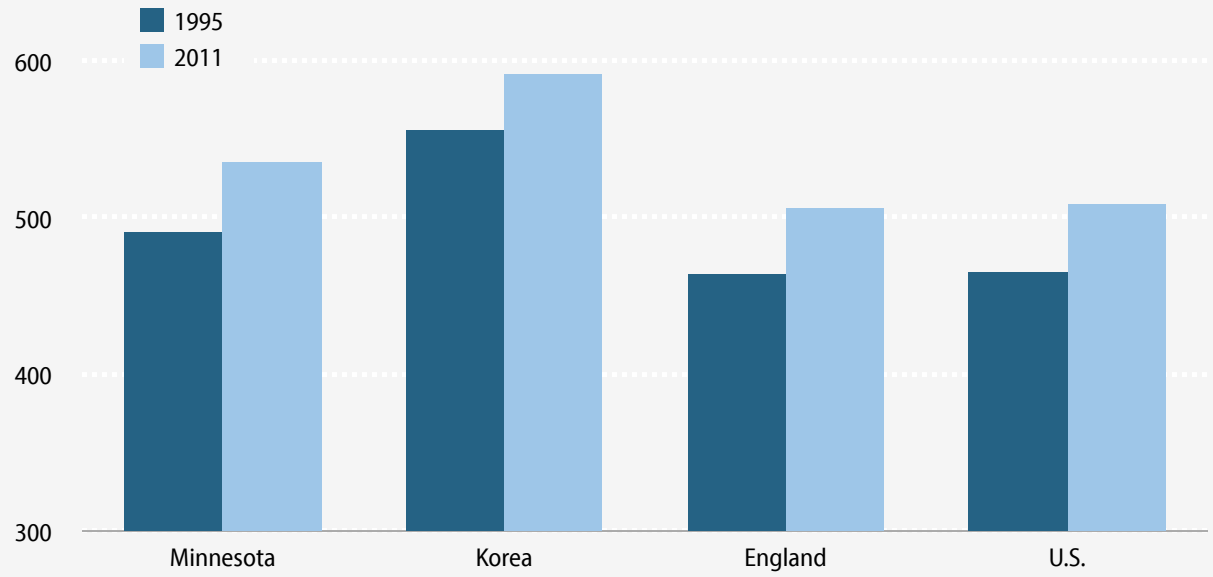


Source: EPI analysis of TIMSS International Database

ECONOMIC POLICY INSTITUTE

FIGURE C

8th grade TIMSS math scores adjusted to 2011 U.S. FAR weights, U.S., Minnesota, Korea, and England, 1995–2011

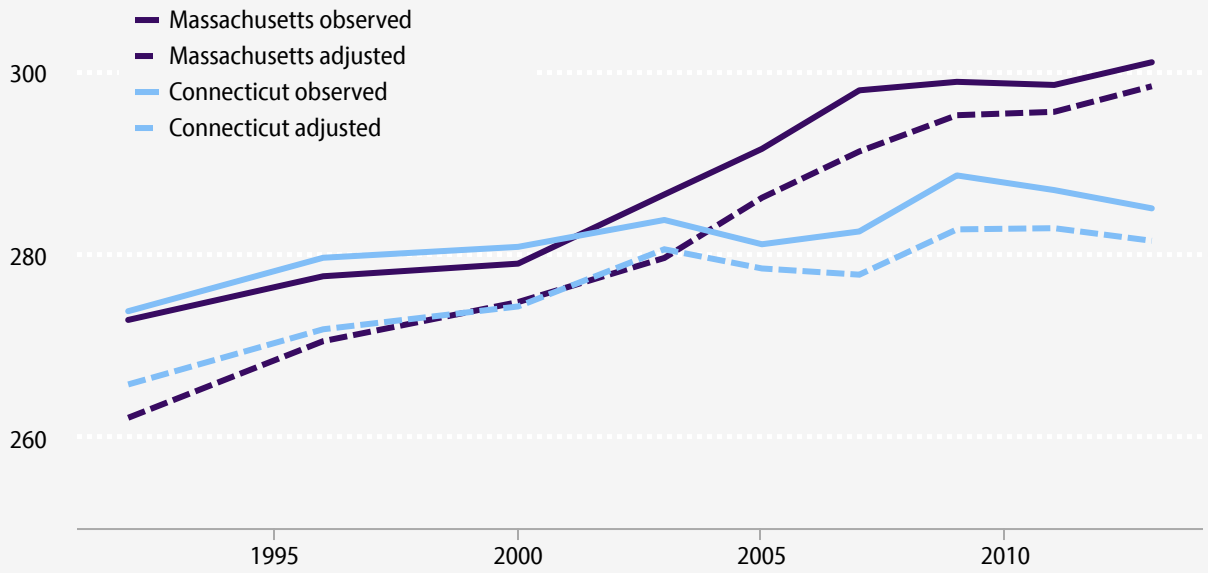


Source: EPI analysis of TIMSS International Database

ECONOMIC POLICY INSTITUTE

FIGURE D

Connecticut and Massachusetts NAEP 8th grade mathematics scores, 1992–2013



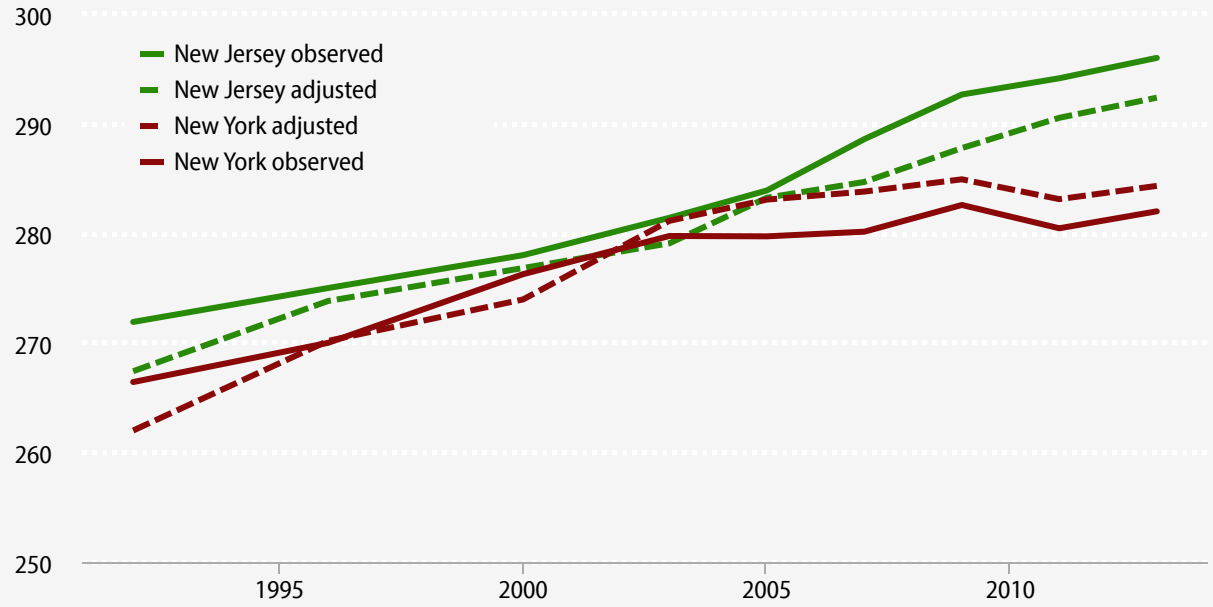
Note: Scores are adjusted within each test year for student characteristics, student family socioeconomic (SES) characteristics, and racial and SES composition of school. Test scores are adjusted across years for changing social class composition of sample by using 1996 student characteristics, family academic resources, and school racial and SES composition to weight all years.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

FIGURE E

New York and New Jersey NAEP 8th grade mathematics scores, 1992–2013



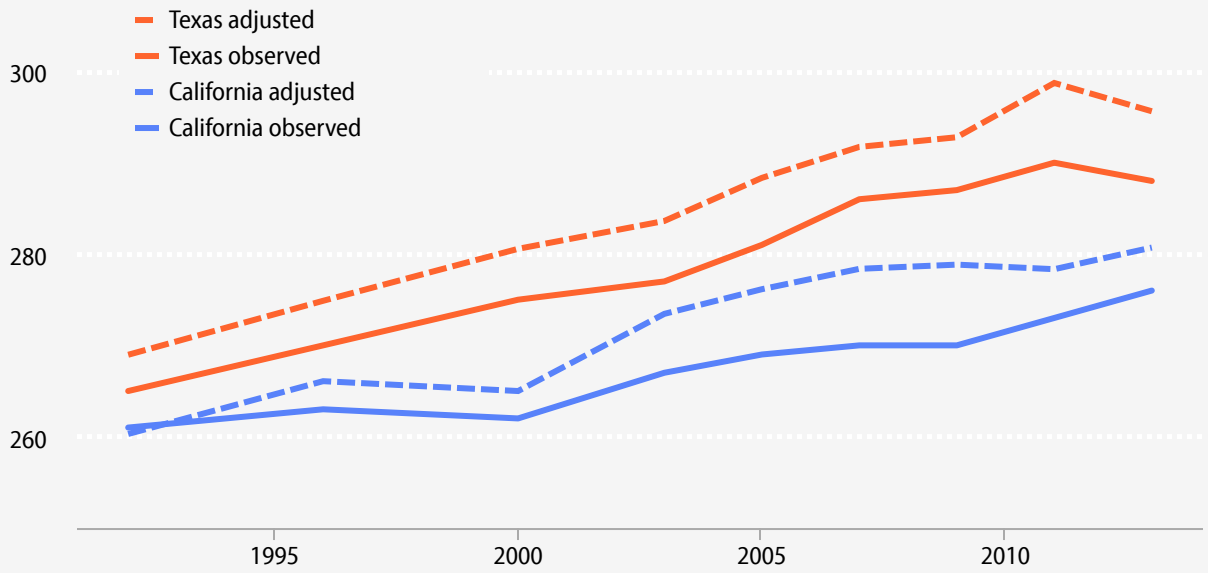
Note: Scores are adjusted within each test year for student characteristics, student family socioeconomic (SES) characteristics, and racial and SES composition of school. Test scores are adjusted across years for changing social class composition of sample by using 1996 student characteristics, family academic resources, and school racial and SES composition to weight all years.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

FIGURE F

California and Texas NAEP 8th grade mathematics scores, 1992–2013



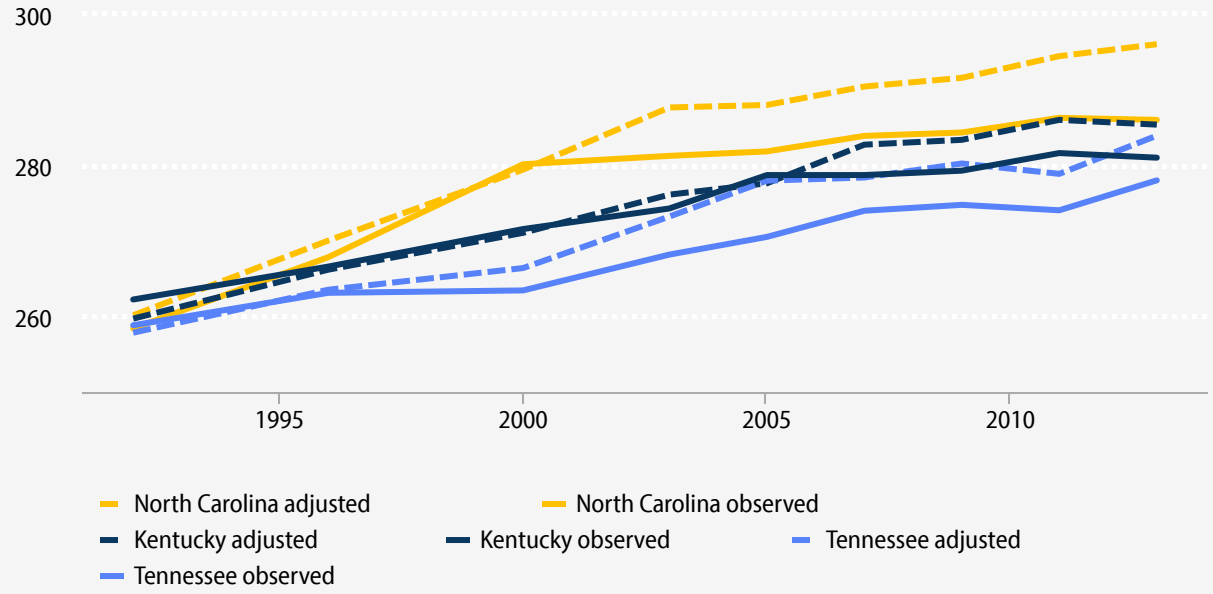
Note: Scores are adjusted within each test year for student characteristics, student family socioeconomic (SES) characteristics, and racial and SES composition of school. Test scores are adjusted across years for changing social class composition of sample by using 1996 student characteristics, family academic resources, and school racial and SES composition to weight all years.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

FIGURE G

North Carolina, Kentucky, and Tennessee NAEP 8th grade mathematics scores, 1992–2013



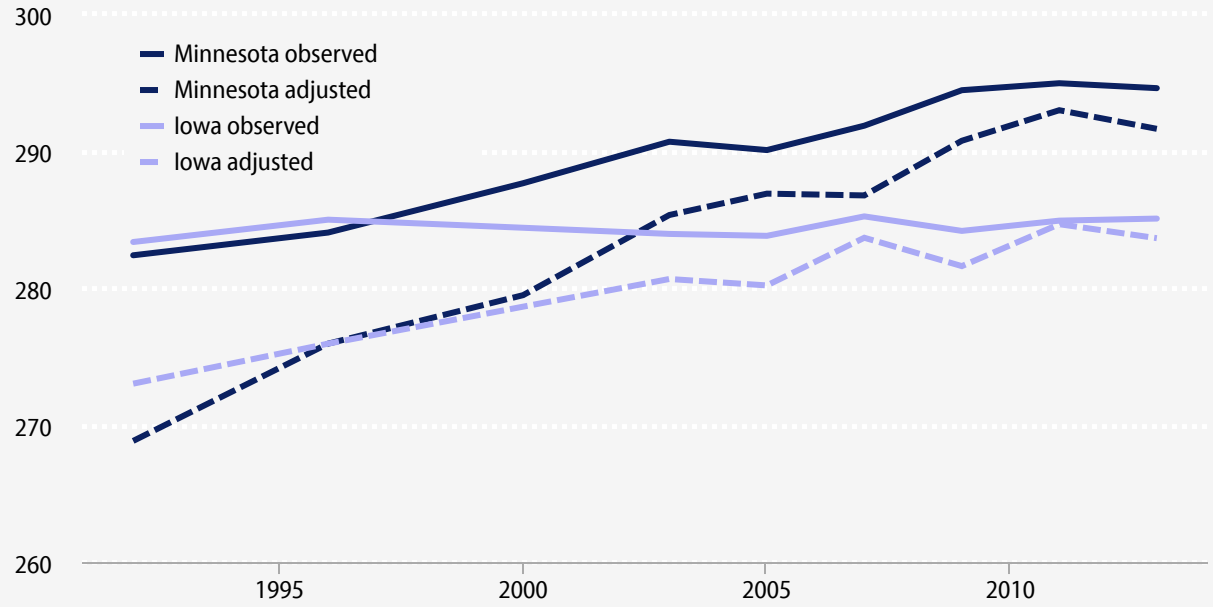
Note: Scores are adjusted within each test year for student characteristics, student family socioeconomic (SES) characteristics, and racial and SES composition of school. Test scores are adjusted across years for changing social class composition of sample by using 1996 student characteristics, family academic resources, and school racial and SES composition to weight all years.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

FIGURE H

Iowa and Minnesota NAEP 8th grade mathematics scores, 1992–2013



Note: Scores are adjusted within each test year for student characteristics, student family socioeconomic (SES) characteristics, and racial and SES composition of school. Test scores are adjusted across years for changing social class composition of sample by using 1996 student characteristics, family academic resources, and school racial and SES composition to weight all years.

Source: EPI analysis of NCES NAEP microdata

ECONOMIC POLICY INSTITUTE

Appendix: Regression analysis of state scores adjusted for FAR measures

In addition to the results in Tables 2A, 2B, 2C, 3A, 3B, 5A, and 5B, we estimated differences in the TIMSS and PISA mathematics performance and PISA reading performance among U.S. states and comparison countries over 1999–2011 (TIMSS) and for 2012 (PISA) using regression analysis. The regressions estimate inter-state and -country differences controlling for a number of student attributes—gender, whether language spoken in the home is the language of the test, whether student age is higher than the average in the sample, books in the home, and mother's education—and the average books in the home of the students in a classroom. Thus, the controls for student FAR are more extensive than those used elsewhere in the paper.

Rows 1, 2, 3, and 4 in **Appendix Table A1** show the regression coefficients for each country and state compared with the U.S. mean score. These differences are estimated using the combined samples of all the countries and states shown. Each regression coefficient in row 1 represents the estimated difference between how well students in that country or U.S. state performed on the TIMSS 1999 mathematics test and how well U.S. students performed when we control for a set of student characteristics and the average books in the home in the student's class (peer effect). The standard errors of the estimated coefficients are in brackets below the coefficient. For estimated performance to be significantly different from the performance of students in the United States as a whole, the coefficient needs to be at least twice as large as the standard error. Thus, students in Canada, Finland, Korea, and England all scored significantly higher than students in the United States on the 1999 TIMSS, and among the four U.S. states that took the 1999 TIMSS, only students in Massachusetts scored significantly higher.

Row 2 of Table A1 shows the estimated regression results for the TIMSS 2011 mathematics test. The adjusted average score for students in the United States increased from 499 to 510, and the difference between student performance in Finland and England compared with the United States was not statistically significant. Students in Alberta performed significantly worse than U.S. students, and students in Quebec and Korea, significantly better. Korean students scored about a standard deviation higher than U.S. students in both 1999 and 2011. Nine states took the TIMSS as separate entities in 2011, and students in many of them scored significantly higher on mathematics than students in the United States as a whole. Students in Massachusetts, Indiana, North Carolina, Minnesota, and Colorado scored significantly higher. Students in Alabama and California scored significantly lower. Students in Connecticut and Florida scored about the same as the United States as a whole. In the states that took the 1999 TIMSS and the 2011 TIMSS, students in Connecticut made an absolute gain adjusted for changes in student characteristics, but no gain relative to students in the United States as a whole. Students in Massachusetts, Indiana, and North Carolina made absolute gains and gains relative to U.S. students as a whole. The relative gains of students in Massachusetts and North Carolina were substantial in this 12-year period.

Although in 1999 students in Massachusetts scored about the same as students in England, students in Canada, Finland, and Korea scored higher in TIMSS mathematics than students in any of the four U.S. states that took the test that year. In 2011, however, students in Massachusetts, North Carolina, and Minnesota scored at least as high as any comparison country or province except Korea. In addition to students in these three states, students in Indiana scored higher than

students in Finland and England, and students in all but Alabama and California scored higher than students in Alberta or Ontario.

We estimated similar regression coefficients for the three states that participated in the PISA test and compared them with our comparison countries, including the United States. The estimated coefficients for those states and comparison countries are shown in Table A1, rows 3 and 4, and the absolute adjusted scores in rows 7 and 8. These results show that students in Massachusetts scored at least as high on the PISA reading as any of the comparison countries, and students in Connecticut scored at least as high as students in all the comparison countries but Korea. Students in Florida scored about the same as students in the United States as a whole and in the U.K., but lower than students in all the other comparison countries.

In the PISA mathematics test adjusted for student characteristics, students in Massachusetts scored at least as high as students in all the comparison countries but Korea, Germany, or Poland. The regression method gives these three high-scoring countries a greater difference over Massachusetts than our adjusted scores using just books in the home in Table 2B. Students in Connecticut only score higher than students in the U.K. and about the same as students in Ireland, and students in Florida score lower or substantially lower than students in all the comparison countries.

APPENDIX TABLE 1

Estimated differences, controlling for measures of family academic resources, in TIMSS and PISA scores between the U.S. national scores and scores in U.S. states, Canadian provinces, and comparison countries

	Canada	Alberta	Ontario	Quebec	Finland	Korea	England	U.S.	Connecticut	Massachusetts	Indiana	North Carolina	Minnesota	Alabama	Colorado	Florida	California
Regression coefficients and U.S. absolute score derived from regressions																	
Mathematics																	
(1) TIMSS 1999	27.34				30.16	91.72	15.97	498.62	7.07	10.3	7.13	-1.73					
	[3.38]				[4.70]	[3.30]	[5.08]		[5.57]	[4.48]	[6.83]	[5.19]					
(2) TIMSS 2011		-7.48	-3.08	29.28	2.51	97.27	4.03	509.65	6.35	43.77	13.95	27.08	29.63	-32.7	8.73	8.69	-15.04
		[3.51]	[2.93]	[3.11]	[3.15]	[2.85]	[4.63]		[3.71]	[5.23]	[4.31]	[5.11]	[4.12]	[4.13]	[3.96]	[4.89]	[4.90]
(3) PISA 2012 math	29.55	26.69	54.29	15.02	36.83	5.08	35.4	14.69	483.65	-11.72	14.35	22.96					
	[3.08]	[3.47]	[4.45]	[3.53]	[3.98]	[3.34]	[3.73]	[2.96]		[5.38]	[4.20]	[4.55]					
Reading																	
(4) PISA 2012	18.29	16.21	22.39	8.99	18.28	-5.32	19.43	20.31	499.62	-2.91	13.68	19.52					
	[3.30]	[4.02]	[4.64]	[3.67]	[3.91]	[4.01]	[4.13]	[3.54]		[5.57]	[4.76]	[5.57]					
Adjusted absolute scores (from regressions)																	
Mathematics																	
(5) TIMSS 1999	525.96				528.78	590.34	514.59	498.62	505.69	508.92	505.75	496.89					
(6) TIMSS 2011		502.17	506.57	538.93	512.16	606.92	513.68	509.65	516	553.42	523.6	536.73	539.28	476.95	518.38	518.34	494.61
(7) PISA 2012 math	513.2	510.34	537.94	498.67	520.48	488.73	519.05	498.34	484	471.93	498	506.61					
Reading																	
(8) PISA 2012 reading	517.91	515.83	522.01	508.61	517.9	494.3	519.05	519.93	500	496.71	513.3	519.14					

Source: EPI analysis of OECD PISA International Database

ECONOMIC POLICY INSTITUTE

Endnotes

1. PISA is sponsored by the Organization for Economic Cooperation and Development (OECD). See http://www.pisa.oecd.org/ and http://nces.ed.gov/surveys/pisa/. PISA was administered in 2000, 2003, 2006, 2009, and 2012.

2. TIMSS was administered by the International Association for the Evaluation of Educational Achievement (IEA) to 8th graders in 1995, 1999, 2003, 2007, and 2011. See http://timss.bc.edu/ and http://nces.ed.gov/timss/. An international test of reading, the Progress in International Reading Literacy Study (PIRLS), was administered only to 4th graders in 2001, 2006, and 2011. TIMSS was also administered to 4th graders simultaneously with the 8th grade administration. We do not discuss 4th grade scores, either from PIRLS or from TIMSS, in this report.

3. For a review of limitations and challenges associated with the utilization of these tests, see Carnoy 2015.

4. The National Assessment of Educational Progress (NAEP) is a nationally representative and continuing assessment of students in various grades and subject areas. It is sponsored by the National Center for Education Statistics of the U.S. Department of Education (http://nces.ed.gov/nationsreportcard). NAEP assessments have been conducted periodically in reading, mathematics, science, writing, U.S. history, civics, geography, and other subjects, beginning in 1969. State assessments are available since 1990.

5. As will be discussed at length in the following section of the paper, this confirms the findings of an earlier EPI report, *What Do International Tests Really Show About American Students' Performance?* (Carnoy and Rothstein 2013). This also applies to the finding presented in the next bullet point.

6. For Connecticut, the difference relative to Korea is more substantial. See details below.

7. Carnoy and Rothstein pointed out that, for example, American adolescents perform relatively well on algebra questions, and relatively poorly on geometry questions, compared with adolescents in other countries. Thus, if a test has more algebra items and fewer geometry items, U.S. students will compare more favorably with students in other countries. Policymakers who draw conclusions about the relative performance of U.S. students from an assessment rarely consider whether there is an appropriate balance between these topics on any particular international assessment. Similar questions arise with regard to a "reading" test. There are undoubtedly subskills covered by international reading and math tests on which some countries are relatively stronger and others are relatively weaker. The report recommended that investigation of these differences should be undertaken before drawing policy conclusions from international test scores.

8. There is no precise way to make family academic resource comparisons between countries (or states). PISA collects data on many characteristics that are arguably related to family academic resource levels, and also assembles them into an overall index (OECD 2013a). Although none of the possible indicators of FAR independently is entirely satisfactory, we think that the number of books in the home is probably superior for purposes of international test score comparisons, and we use it to divide students into FAR groups. A very high fraction of students in both the PISA and TIMSS surveys answer the question concerning books in the home, something less true for other important FAR indicator questions asked on the student questionnaires. An alternative indicator would be student reported mother's education, which is highly correlated in the countries and states using PISA and TIMSS data. In the NAEP data we do not use books in the home because the categories of this variable in the NAEP are very broad.

9. See more details about family academic resources in endnote 8. In the Appendix we use regression analysis to adjust scores by family academic resource levels, and there we use both mother's education and books in the home. The results in the Appendix show that the differences in the two ways of adjusting scores are generally very small. They also support the notion that using

books in the home as a single measure of family academic resources provides reasonably accurate adjustments for FAR differences among students.

10. U.S. FAR Group 1 students score higher in math and reading than FAR Group 1 students in France, Germany and the U.K. They score lower than FAR Group 1 students in Canada, Korea, Finland, Poland and also slightly below Ireland, but the gaps are smaller than the gaps between higher FAR groups in the United States relative to high FAR groups in other countries. Others have made the same argument, using the 2012 PISA results, showing that students in the top socioeconomic decile in the United States *taken as a whole* scored below students in the top socioeconomic decile in many countries. See Diehm and Resmovits 2014.

11. Upper-middle and higher FAR groups in Massachusetts and Connecticut score about the same or at least as well in reading as comparable students in all the other higher-scoring countries (including Korea). In mathematics, students in Massachusetts in Groups 4–6 (upper-middle, higher, and highest FAR) also score higher than students in Canada, Finland, Poland, and Ireland. Indeed, when adjusted for differences in the FAR distributions across countries, adjusted absolute scores presented in Appendix Table A1 show that all students in Massachusetts score at least as well as students in every one of our comparison countries except in Korea in reading using PISA 2012 and mathematics using TIMSS 2011 (in the same table, using the PISA 2012 mathematics test, adjusted scores for Massachusetts are slightly lower than they are in some of the comparison countries).

12. We can only speculate why advantaged students in Florida do not have correspondingly higher scores on the PISA in reading and mathematics as do advantaged students in Massachusetts and Connecticut or in France, Germany, and the U.K. One possible explanation is that Florida public schools have put more emphasis on raising achievement among lower FAR students relative to higher FAR students, and the opposite is true in Connecticut and Massachusetts (for Florida, see Figlio and Rouse 2006). A second possible explanation is that teachers in Florida are of high enough "quality" to raise achievement among lower FAR students achieving at lower levels but not of high enough quality to raise the achievement of advantaged students. In contrast, teachers in Massachusetts and Connecticut may be of much higher "quality"—high enough to push high-achieving students to even higher levels.

13. We deemphasize comparisons with Group 6 students because the PISA results for U.S. Group 6 students behave so erratically in 2012 compared with previous years and compared with Group 5 student scores. It is possible that many students in the U.S. sample misreported that they had more than 500 books in the home, when, in fact, they had many fewer books in the home. This would have resulted in substantially underestimated average scores for Group 6 in 2012.

14. We have student performance on PISA in three U.S. states but only for one year, 2012. Thus, we cannot compare changes in student performance across states on the PISA test over time.

15. Carnoy and Rothstein (2013) showed that TIMSS and NAEP scores in these states move in similar directions and that the gains over time are also similar, particularly over the longer period of 1995/1999 to 2011. However, North Carolina makes a larger gain on the TIMSS (0.7 percent annually) than on the NAEP (0.3 percent annually). In 2011, the TIMSS and the NAEP were aligned, and the National Center for Education Statistics (NCES), which applies the NAEP, was able to make estimates of 2011 TIMSS scores in mathematics and science for every U.S. state, based on the results of the nine "linking states" that actually participated in the TIMSS (NCES 2013).

16. As reported by Stewart (2014), PISA Director Andreas Schleicher conceded that the Shanghai sample only represented 73 percent of that province's 15-year-olds during his address to the British House of Commons Education Select Committee.

17. Some have posed doubts as to whether higher scores may be at least partly due to the massive amount of out-of-school tutoring and test prep engaged in by East Asian students.

18. There is a vast literature on cram school in Korea (*hagwon*), Japan (*juku*), and other Asian countries, and here we only cite a few references. However, there is no doubt that a high percentage of students in these countries spend a considerable amount of time during their middle school and high school years in cram schools/courses in addition to studying for tests and completing other work for "regular" school. Families invest major resources in extra instruction. Surprisingly, this is rarely mentioned when discussing whether such behavior or levels of investment are broadly transferable to other societies.

19. Further, in some high-scoring countries, the results on the PISA or the TIMSS tests are considered a matter of national "legitimacy," on par with performing well in international sports events. PISA proctors may exhort students picked to take the test to make a supreme effort for the "nation" (Carnoy et al. 2014). In other countries, including the United States, the PISA or TIMSS is just another low-stakes test that students are asked to take. Such confounding contextual factors make it much more difficult for the United States to extract education policy lessons from comparisons with other countries.

20. For example, if we randomly assigned Japanese teachers and mathematics curriculum to U.S. classrooms, how much better would students in those classrooms fare than students matched with U.S. mathematics teachers?

21. As explained by Dillon (2010), when testifying before a U.S. congressional committee considering the reauthorization of the Elementary and Secondary Education Act, PISA Director Andreas Schleicher said Finland had the world's "best performing education system."

22. The PISA test may not be the most relevant measure of performance and progress in mathematics. As we have shown (Table 2B), the United States performs much lower than Finland in mathematics on the PISA test even when we adjust scores for FAR differences in the two countries' PISA student samples. But the opposite is the case when we compare FAR-adjusted Finnish and U.S. mathematics performance on the 2011 TIMSS. Average adjusted U.S. scores were higher than in Finland in 2011 (Table 3A).

23. It is also possible that Poland, like Estonia (Carnoy et al. 2014) and a number of other countries, considers improved performance on the PISA test as a means of achieving greater "legitimacy" in the international community, and has gradually reformed its curriculum and internal student evaluation instruments to match the types of questions asked on the PISA test. As students become more familiar with such questions, it would not be surprising that they would perform better.

24. Germany is also a federal system in which the bulk of the responsibility for managing the education systems and implementing reforms lies in the *lande*, the equivalent of German states. Germany, like Australia, Brazil, and Mexico (among others), has regularly applied very large national PISA samples in order to obtain large enough random samples to analyze differences among states. We know from early PISA tests (2003) that the differences among German states, as in Brazil, Mexico, and the few U.S. states that took the PISA, are substantial (Woessmann 2007). However, for political reasons—mainly sensitivity to comparisons with the former East and West German states—Germany has not allowed analysis of state differences for later tests using the full state randomized PISA data. Like our analysis of the U.S. state differences, exploring differences among German states would undoubtedly show that some have made significantly larger gains than others. The lessons learned from that analysis would undoubtedly be more useful than those contained in *Lessons from PISA for the United States*.

25. As noted previously, NAEP student data at the state level are not available for every state over 1992–2000. Since 2003, it is mandatory for all states to participate in the NAEP, and therefore, we have student-level data in all states over 2003–2013, but not for all states over the full 1992–2013 period.

26. Because of the erratic behavior of the estimated state fixed effect for Alaska, we did not report the Alaska coefficients, although the Alaska sample was included in our regressions.

27. To calculate a test score value for the California reference variable, we estimated the California test score at the mean of the control variables—these included student characteristics, school demographic composition, and teacher characteristics. In Table 7, we estimated the California test score at the mean of the 1996 controls for student characteristics and school composition in order to adjust the average state test scores for changes over time in the demographic composition of the state samples.

28. We also estimated rankings based on controlling for only the first two sets of variables (excluding teacher variables). The state rankings are essentially the same, and the range of scores between the highest- and lowest-scoring states are also essentially the same. Thus, adding controls for the limited set of teacher characteristics variables available in the NAEP to the controls for student family characteristics and school poverty and race concentration contributes very little to our understanding of why test scores vary among states (see Appendix). This is consistent with much of the production literature in the United States concerning the impact of teacher characteristics on student achievement. Although teacher experience has a significant effect on student achievement, it tends to be small (see, for example, Clotfelter et al. 2007; Goldhaber and Brewer 2000).

29. In Table 7, we show the adjusted scores including only the first two sets of variables for 8th grade mathematics for 1992–2013.

30. The adjustment reduces the variance of average 8th grade mathematics scores among states by 76 percent in 2003, 66 percent in 2005, 62 percent in 2007, 65 percent in 2009, 59 percent in 2011, and 60 percent in 2013. The reported NAEP 8th grade math scores had the following interstate variance: 76.87 in 2003, 73.47 in 2005, 75.91 in 2007, 73.04 in 2009, 57.73 in 2011, and 52.28 in 2013. The variances of the adjusted scores were 18.56 in 2003, 24.82 in 2005, 28.62 in 2007, 25.53 in 2009, 23.45 in 2011, and 21.00 in 2013.

31. We explored the possibility that 8th grade adjusted scores in math and reading across states were significantly correlated with 4th grade adjusted scores four years earlier—a grade cohort effect. We tested 2003 4th grade against 2007 8th grade, 2005 4th grade against 2009 8th grade, 2007 4th grade against 2011 8th grade, and 2009 4th grade against 2013 8th grade. Although the 4th grade scores in 2003, 2005, 2007, and 2009 are highly correlated with the 8th grade scores four years later, the 4th grade scores in 2003, 2005, 2007, and 2009 are even more highly correlated with 8th grade scores in the same year. We concluded that more than a cohort effect, 4th and 8th grade scores tend to move together in states in the same year. Results of these regressions are available from the authors on request.

32. We were not able to include teacher characteristics in the adjustment because the definition of these variables changed somewhat in the 1992, 1996, and 2000 versions of the NAEP. Nevertheless, the difference in adjusted scores when the teacher variables are included or excluded over 2003–2013 (when we can compare the two adjustments) is very small.

33. The scores in Table 7 are also adjusted for changes in average student and school demographics across years. Within-year adjusted state test scores are estimated as if the student and school demographic composition were the same as in 1996. We chose 1996, and not the base year, 1992, because in 1992, we did not have data on individual free and reduced lunch status for each student.

34. States such as Connecticut have relatively higher reading gains than mathematics gains; students in Texas have made very large gains in 8th grade mathematics compared with students in other states, but more modest gains in 4th grade mathematics and in the NAEP reading test, both in the 8th and 4th grades.

35. As in Tables 6A1, 6A2, 6B1, and 6B2, we have omitted the Alaska results because of erratic variation in adjusted test scores in the state.

36. The measure is the Collective Bargaining Coverage (CBC) rate from the CPS survey and measures the proportion of teachers covered by collective bargaining (as opposed to a measure of union membership). The sample used to construct this variable is

from the CPS-ORG data on full-time public K–12 school teachers with at least a bachelor's degree and with imputed or non-imputed positive weekly earnings, for a pooled sample from 2009 to 2013 (from Garcia and Mishel, forthcoming).

37. To account for the multi-level nature of the analysis (time, state), we "cluster adjust" the standard error estimates on the states. Regression results are available from the authors upon request.

38. From the NAEP reported results (NCES, NAEP data tool), and based on average tests scores by ethnic groups over time, the impact was apparently relatively greatest on black and Hispanic students, since white students in Connecticut scored about the same as white students in Massachusetts across FAR groups.

39. See MSU 2008.

References

Bray, M. 2006. "Private Supplementary Tutoring: Comparative Perspectives on Patterns and Implications." *Compare*, vol. 36, no. 4, 515–530.

Bray, M., and Lykins, C. 2012. *Shadow Education*. Manila: Asian Development Bank.

Byun, Soo-yong. 2014. "Shadow Education and Academic Success in Republic of Korea." In H. Park & K. Kim (editors), *Korean Education in Changing Economic and Demographic Contexts* (39–58). Singapore: Springer.

Carnoy, Martin. 2015, forthcoming. *International Test Score Comparisons and Educational Policy: A Review of the Critiques*. National Education Policy Center at the University of Colorado.

Carnoy, Martin, and Susanna Loeb. 2003. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Educational Evaluation and Policy Analysis,* vol. 24, no. 4.

Carnoy, Martin, and Richard Rothstein. 2013. *What Do International Tests Really Show About American Student Performance?* Economic Policy Institute.

Carnoy, Martin, Tatiana Khavenson, and Alina Tatiana. 2014. "Using TIMSS and PISA Results to Inform Educational Policy: A Study of Russia and Its Neighbors." *Compare*, vol. 45, no. 2, 248–271.

Carnoy, Martin, Tatiana Khavenson, Prashant Loyalka, William Schmidt, and Andrey Zakharov. 2015a. "Revisiting the Relationship Between International Assessment Outcomes and Educational Production: Evidence from a Longitudinal PISA-TIMSS Sample." Graduate School of Education, Stanford University.

Carnoy, Martin, Tatiana Khavenson, Leandro Costa, Izabel Fonseca, and Luana Marotta. 2015b. "Is Brazilian Education Improving? A Comparative Foray Using PISA and SAEB Brazil Test Scores." Stanford Graduate School of Education.

Clotfelter, Charles T., Helen Ladd, and Jacob Vigdor. 2007. "Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects." *Economics of Education Review,* vol. 26, no. 6, 673–682.

Diehm, Jan, and Joy Resmovits. 2014. "Surprising Test Results for Some of the World's Richest Students." *Huffington Post,* January 23.

Dillon, Sam. 2010. "Many Nations Passing U.S. in Education, Expert Says." *New York Times*, March 10.

Duncan, Arne. 2012. "Statement by U.S. Secretary of Education Arne Duncan on the Release of the 2011 TIMSS and PIRLS Assessments." U.S. Department of Education, December 11.

Duncan, Arne. 2013. "The Threat of Educational Stagnation and Complacency: Remarks of U.S. Secretary of Education Arne Duncan at the Release of the 2012 Program for International Student Assessment (PISA)." U.S. Department of Education, December 3.

Figlio, David, and Cecilia Rouse. 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics*, vol. 90, nos. 1–2, 239–255.

García, Emma, and Lawrence Mishel. Forthcoming. *Unions and the Allocation of Teachers in Public Schools*. Economic Policy Institute.

Goldhaber, D.D., and D.J. Brewer. 2000. "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement." *Education Evaluation and Policy Analysis*, vol. 22, 129–145.

Hanushek, E. 1986. "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature*, vol. 24, no. 3, 1141–1177.

Hanushek, E.A., and M.E. Raymond. 2001. "The Confusing World of Educational Accountability." *National Tax Journal*, vol. 54, 365–384.

Harvey, James. 2015. "Ten Things You Need to Know About International Assessments." *Washington Post*, February 3.

Hiebert, J., R. Gallimore, H. Garnier, K. Givvin, H. Hollingsworth, J. Jacobs, A.M. Chui, D. Wearne, M. Smith, N. Kersting, A. Manaster, E. Tseng, W. Etterbeek, C. Manaster, P. Gonzales, and J. Stigler. 2003. *Teaching Mathematics in Seven Countries: Results from the TIMSS 1999 Video Study.* U.S. Department of Education National Center for Education Statistics.

Hoxby, Caroline. 1996. "How Teachers Unions Affect Education Production." *Quarterly Journal of Economics,* vol. 111, no. 3, 671–718.

Jakubowski, Maciej, Harry Patrinos, Emilio Ernesto Porta, and Jerzy Wisniewski. 2010. *The Impact of the 1999 Education Reform in Poland*. World Bank Policy Research, Working Paper No. 5263.

Klein, Ruben. 2011. "A Reanalysis of PISA Results: Comparability Problems." *Ensaio: Avaliação e Políticas Públicas em Educação*, vol. 19, no. 73, October/December.

Lee, J., and K.K. Wong. 2004. "The Impact of Accountability on Racial and Socioeconomic Equity: Considering Both School Resources and Achievement Outcomes." *American Educational Research Journal,* vol. 41, 797–832.

Loveless, Tom. 2013. "PISA's China Problem." Brookings Institution, *Brown Center Chalkboard,* October 9.

Loveless, Tom. 2014. "Lessons from the PISA-Shanghai Controversy." Brookings Institution, *Brown Center Chalkboard,* March 18.

Loyalka, P., E. Kardanova, L. Liu, V. Novdonov, H. Shi, K. Enchicova, N. Johnson, and Liyang Mao. 2015. *Where Are the Skilled Engineers Coming From? Assessing and Comparing Skill Levels and Gains in Engineering Programs Across the US, China, and Russia*. Stanford University, working paper.

Lubienski, Sarah Theule. 2002. "A Closer Look at Black-White Mathematics Gaps: Intersections of Race and SES in NAEP Achievement and Instructional Practices Data." *The Journal of Negro Education*, vol. 71, no. 4, 269–287.

Lubienski, Christopher, and Sarah Theule Lubienski. 2014. *The Public School Advantage: Why Public Schools Outperform Private Schools*. Chicago: University of Chicago Press.

Matloff, Norman. 2013. *Are Foreign Students the 'Best and Brightest'?* Economic Policy Institute Briefing Paper.

Medrich, Elliott, and Jeanne Griffith. 1992. *International Mathematics and Science Assessment: What Have We Learned?* U.S. Department of Education, National Center for Educational Statistics, Office of Educational Research and Improvement.

Michigan State University (MSU). 2008. "MSU Scholars Help Minnesota Become Global Leader in Math." Press release, December 9.

Murnane, Richard, John Willett, and Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wages." *Review of Economics and Statistics*, vol. 77, no. 2, 251–266.

National Center for Education Statistics (NCES). 2013. *The Nation's Report Card: U.S. States in a Global Context: Results From the 2011 NAEP-TIMSS Linking Study.*

National Center for Education Statistics (NCES). Various years. NCES NAEP microdata [unpublished data].

Organization for Economic Cooperation and Development (OECD) Program for International Student Assessment (PISA). Various years. *PISA International Database*.

Organization for Economic Cooperation and Development (OECD). 2011. *Strong Performers and Successful Reformers in Education: Lessons From PISA for the United States*.

Organization for Economic Cooperation and Development (OECD). 2013a. *PISA 2012 Results: What Students Know and Can Do: Student Performance in Mathematics, Reading and Science (Volume I).*

Organization for Economic Cooperation and Development (OECD). 2013b. *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II).*

Organization for Economic Cooperation and Development (OECD). 2013c. *PISA 2012 Results: What Makes Schools Successful? Resources, Policies, and Practices (Volume IV).*

Organization for Economic Cooperation and Development (OECD). 2013d. *Strong Performers and Successful Reformers in Education: Lessons From PISA for the United States*.

Ravitch, Diane. 2013. "My View of the PISA Scores." *Diane Ravitch's Blog,* December 3.

Ripley, A. 2013. *The Smartest Kids in the World.* New York*:* Simon and Shuster.

Schmidt, William H., Curtis C. McKnight, and Senta A. Raizen. 1997. *A Splintered Vision: An Investigation of U.S. Science and Mathematics Education.* Dordrecht, The Netherlands: Kluwer.

Schmidt, W.H., C. McKnight, R. Houang, H. Wang, D. Wiley. 2001. *Why Schools Matter: A Cross-National Comparison of Curriculum and Learning.* San Francisco: Jossey-Bass.

Stanat, P., D. Rauch, and M. Segeritz. 2010. "Schülerinnen und Schüler mit Migrationshintergrund." In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, and P. Stanat, editors. *PISA 2009. Bilanz nach einem Jahrzehnt.* Münster, Germany: Waxmann, 200–230.

Stewart, William. 2013. "Is PISA Fundamentally Flawed?" *Times Education Supplement Magazine,* July 26, updated September 27, 2014.

Stewart, William. 2014. "More Than a Quarter of Shanghai Pupils Missed by International PISA Rankings." *Times Education Supplement*, March 6.

Trends in International Mathematics and Science Study (TIMSS). Various years. *TIMSS International Database*.

Wantanabe, Manabu. 2013. *Juku: The Stealth Force of Education and the Deterioration of Schools in Japan.* Independently published.

Woessmann, L. 2007. *Fundamental Determinants of School Efficiency and Equity: German States as a Microcosm for OECD Countries*. CESifo Working Paper 1981.